

La ridondanza costitutiva

Rappresentazioni multiple, identità del contenuto e la struttura distribuita della mente

Capitolo I — Il paradosso della molteplicità: come la stessa cosa abita luoghi diversi nella mente

I.1 — Un fatto scomodo al centro della neuroscienza cognitiva

Esiste una tensione sotterranea, raramente tematizzata con sufficiente rigore, tra il modo in cui la filosofia del contenuto mentale concepisce le rappresentazioni e il modo in cui le neuroscienze cognitive le osservano. La filosofia — da Fodor a Millikan, da Dretske alle teorie predittive che hanno dominato il decennio scorso — assume, quasi invariabilmente, che il contenuto di uno stato mentale sia localizzabile: vi sarebbe un simbolo, un nodo funzionale, una distribuzione di probabilità a posteriori che è la rappresentazione di un dato oggetto o proprietà. Questa localizzabilità è così profondamente radicata nelle assunzioni di sfondo del dibattito che raramente viene tematizzata come presupposizione: essa appare, piuttosto, come la condizione ovvia di possibilità di qualsiasi teoria del contenuto che aspiri a un grado minimo di determinatezza. Se il contenuto non è localizzabile, se non vi è alcun vettore o simbolo o distribuzione che lo *porta*, allora di cosa stiamo parlando quando parliamo di rappresentazioni mentali?

La neuroscienza, nel frattempo, mostra con un'insistenza che ha ormai il carattere dell'ovvietà empirica che nessun contenuto risiede in un unico luogo. Quando un essere umano percepisce un volto familiare, tale volto è simultaneamente rappresentato nella corteccia visiva primaria, dove le proprietà di basso livello come orientamento dei bordi, frequenza spaziale e contrasto locale sono codificate in modo retinotopico; nel giro fusiforme anteriore, dove l'identità del volto emerge come pattern globale di risposta di una popolazione neurale; nel solco temporale superiore, sensibile al movimento del volto e alle sue componenti

espressive; nel cortex entorinale e nell'ippocampo, dove l'identità del volto si lega a una rete di associazioni biografiche e spaziali; nell'amigdala, dove la valenza emotiva del volto è codificata come modulazione del tono di eccitazione; nelle regioni prefrontali mediali, dove le aspettative sociali e il modello della persona conosciuta sono mantenuti come strutture predittive. Ciascuna di queste strutture mantiene una codifica distinta — per orientamento, per identità, per valenza emotiva, per contesto episodico, per aspettativa sociale — di quello che, da un punto di vista intuitivo, è il medesimo oggetto.

Questo fatto — che chiameremo, con termine tecnico ma provvisorio, ridondanza rappresentazionale — non è considerato problematico dalla neuroscienza nel senso forte del termine: la neuroscienza non ha, in senso stretto, bisogno di un concetto di identità del contenuto per procedere nelle sue indagini. Le singole aree vengono studiate per le loro funzioni specifiche, i loro profili di risposta, le loro connessioni anatomiche e funzionali. Che tutte queste aree rappresentino «lo stesso volto» è un'affermazione che la neuroscienza presuppone senza mai fermarsi a giustificarla teoricamente, perché la giustificazione appartiene a un livello di analisi — quello del contenuto — che la neuroscienza non ha i propri strumenti per indagare.

La filosofia, invece, ha tutto il motivo di essere turbata da questo stato di cose. La domanda che la ridondanza pone è esattamente quella che qualsiasi teoria del contenuto deve affrontare: se la stessa cosa è rappresentata in luoghi diversi, da rappresentazioni con strutture interne diverse, codifiche diverse, dinamiche di aggiornamento diverse e substrati neurali non sovrapposti, che cosa rende tutte queste rappresentazioni rappresentazioni *della medesima cosa*? E, rovesciando il problema: se le neuroscienze ci dicono che quelle rappresentazioni *sono* della stessa cosa — e lo dicono, implicitamente, ogni volta che parlano di «la rappresentazione del volto» o «la codifica del concetto» — cosa ci dice questa affermazione sulla natura del contenuto?

Il paradosso affiora in tutta la sua profondità non appena si esaminano le opzioni disponibili. La prima opzione è sostenere che solo una delle rappresentazioni multiple sia *la vera* rappresentazione del contenuto, e le altre siano derivate, ausiliarie, copie prive di contenuto proprio. Questa posizione ha una certa attrattiva teorica perché preserva la localizzabilità del contenuto: vi è ancora un luogo in cui il contenuto risiede, anche se è circondata da strutture ancillari che ne amplificano, distribuiscono o traducono il segnale. Ma questa risposta è empiricamente insostenibile in modo piuttosto diretto: se si danneggia il giro fusiforme anteriore, la capacità di riconoscere i volti come tali è gravemente compromessa anche quando le altre strutture sono intatte, producendo il quadro clinico della prosopagnosia associativa; se si danneggia l'ippocampo, i volti non vengono più collocati in un contesto biografico pur essendo ancora riconoscibili come categorie di facce, producendo la dissociazione caratteristica del syndrome amnesico; se si danneggia l'amigdala, il volto è

riconoscibile e contestualizzabile ma perde la sua carica affettiva, producendo quadri di disconoscimento emotivo che possono assumere la forma di delirio di Capgras. Nessuna delle strutture è eliminabile senza perdita funzionale specifica, e la perdita è ogni volta qualitativamente diversa, il che suggerisce che ciascuna contribuisce in modo irriducibile a un aspetto distinto del contenuto complessivo. Sostenere che una sola di esse sia la «vera» rappresentazione equivale a scegliere arbitrariamente quale aspetto del contenuto meriti l'etichetta di «contenuto vero».

La seconda opzione, simmetricamente opposta, è sostenere che ognuna delle strutture ha il suo contenuto proprio e distinto — il giro fusiforme rappresenta *l'identità del volto*, l'amigdala rappresenta *la valenza emotiva del volto*, l'ippocampo rappresenta *il contesto episodico del volto* — e che ciò che chiamiamo «la rappresentazione del volto» è in realtà un aggregato di contenuti distinti che coesistono nel medesimo sistema cognitivo senza costituire un'unità vera e propria. Questa risposta ha il vantaggio di essere fedele all'eterogeneità delle codifiche osservate neurobiologicamente, ed è implicita in una tradizione di modularismo cognitivo che va da Fodor a Coltheart. Ma paga un prezzo teorico altissimo: dissolve l'unità del contenuto in una molteplicità di micro-contenuti, riproducendo a scala neurale il problema del binding che la codifica predittiva pretendeva di aver risolto, e soprattutto lascia aperta — anzi, riapre con urgenza — la domanda in virtù di che cosa questi micro-contenuti distinti costituiscano insieme la rappresentazione di un unico oggetto. Se il giro fusiforme rappresenta un'identità facciale, l'amigdala una valenza emotiva e l'ippocampo un contesto biografico, in che senso questi tre contenuti distinti sono *del medesimo volto*? La domanda dell'unità del contenuto non è eliminata ma spostata di un livello: non dobbiamo più spiegare come una singola rappresentazione porta il contenuto di un volto completo, ma dobbiamo spiegare come tre micro-contenuti distinti si coordinano per costituire il contenuto integrato di quel volto.

L'ipotesi che si svilupperà in questa monografia è che entrambe le risposte siano sbagliate per la stessa ragione di fondo: presuppongono che il contenuto sia una proprietà di singole rappresentazioni, una caratteristica che ciascuna rappresentazione possiede intrinsecamente e indipendentemente dalle altre. La risposta localista attribuisce il contenuto a una singola rappresentazione privilegiata; la risposta frammentata attribuisce micro-contenuti a ciascuna delle rappresentazioni considerate individualmente. L'assunzione di fondo — che il contenuto sia una proprietà che le rappresentazioni portano individualmente — è condivisa da entrambe le posizioni, e la presente monografia argomenterà che è precisamente quell'assunzione a essere erranea.

La tesi opposta — che si argomenterà con strumenti tratti dalla teoria dell'informazione parziale, dalla topologia algebrica e dalla filosofia del linguaggio — è che il contenuto sia una proprietà relazionale che emerge

dalla struttura di coerenza tra rappresentazioni multiple. Non vi è un contenuto che poi si distribuisce in molte sedi; vi sono molte rappresentazioni la cui struttura di relazione reciproca *costituisce* il contenuto. La ridondanza non è un fatto empirico accidentale che una buona teoria del contenuto deve saper accomodare: è la struttura necessaria mediante la quale il contenuto stesso esiste. Chiamiamo questa tesi la tesi della ridondanza costitutiva, non soltanto descrittiva o strumentale, per sottolineare che la molteplicità non è una conseguenza del fatto che il contenuto è presente — una forma di ridondanza puramente esecutiva, per robustezza o efficienza — ma è la condizione di possibilità dell'esistenza stessa del contenuto.

Questa affermazione richiede una elaborazione e una difesa considerevoli, che occuperanno l'intero volume. Ma prima di procedere a quella elaborazione, conviene chiarire la portata della domanda, delimitarne il campo di applicazione, e distinguerla da questioni superficialmente simili che hanno già ricevuto risposte soddisfacenti nel dibattito.

I.2 — La ridondanza non è ciò che sembra: una distinzione tassonomica necessaria

Prima di procedere con la parte costruttiva dell'argomentazione, è necessario sgombrare il campo da un equivoco terminologico che rischia di offuscare l'intera analisi successiva. Il termine «ridondanza» proviene dalla teoria dell'informazione di Shannon, dove ha un significato tecnico preciso e parzialmente diverso dall'uso che se ne farà nel corso di questa ricerca. Comprendere sia il significato originale sia le ragioni del suo scarto rispetto all'uso presente non è un esercizio di igiene terminologica fine a se stesso: è necessario per capire in che senso la ridondanza di cui si tratta qui sia una proprietà più profonda e filosoficamente più rilevante della ridondanza nel senso di Shannon.

In Shannon, la ridondanza di un sistema di codifica è definita come la differenza tra la capacità massima del canale e la quantità di informazione effettivamente trasmessa per simbolo. Formalmente, se $H_{\max}$ è l'entropia massima possibile per un alfabeto di simboli dato e H_{eff} è l'entropia effettiva del messaggio, la ridondanza relativa è:

$$R = 1 - H_{\text{eff}} / H_{\max}$$

Un sistema è completamente ridondante quando $H_{\text{eff}} = 0$, cioè quando i simboli trasmessi non portano alcuna informazione nuova perché il ricevente li può predire con certezza assoluta dal contesto; è completamente privo di ridondanza quando $H_{\text{eff}} = H_{\max}$, cioè quando ogni simbolo è completamente imprevedibile dal contesto. Le lingue naturali hanno ridondanze dell'ordine del 50–75%: in un testo scritto in italiano, la conoscenza delle lettere precedenti permette di

predire con probabilità notevolmente superiore al caso la lettera successiva, il che significa che gran parte dell'informazione apparente è già contenuta nel contesto.

Questa forma di ridondanza — chiamiamola ridondanza codificatoria, o ridondanza nel senso di Shannon — non è quella rilevante per la presente indagine, o non lo è primariamente. La ridondanza codificatoria è una proprietà di un segnale in relazione alla capacità del canale che lo trasmette; è un fenomeno di livello algoritmico, nel senso della tassonomia di Marr, che riguarda l'efficienza della codifica informazionale ma non la struttura del contenuto. Nella terminologia di Marr, la ridondanza codificatoria è un fenomeno del secondo livello — quello algoritmico — e non dice nulla di diretto sulla natura di ciò che viene codificato. È indifferente, in linea di principio, al contenuto semantico: un codice per trasmettere messaggi meningei e un codice per trasmettere comandi motori possono avere la stessa ridondanza codificatoria senza che ciò implichi alcuna relazione tra i rispettivi contenuti.

Quello che interessa qui è un fenomeno diverso, che si potrebbe chiamare ridondanza rappresentazionale o, per anticipare la terminologia che si svilupperà: ridondanza costitutiva. Si tratta della coesistenza simultanea, nel medesimo sistema cognitivo, di molteplici rappresentazioni che in qualche senso concernono il medesimo oggetto o proprietà del mondo. Questa ridondanza non è un'inefficienza della codifica in senso shannoniano: è una caratteristica strutturale dell'architettura cognitiva che, come si argomenterà, ha ragioni funzionali profonde e implicazioni filosofiche di larga portata.

La distinzione può essere resa più precisa attraverso un esempio. Supponiamo che il sistema visivo codifichi la percezione di un quadrato rosso attraverso una rappresentazione neurale distribuita: alcune cellule sono attive per il colore rosso, altre per la forma quadrata, altre per la combinazione di colore e forma, altre ancora per la collocazione nello spazio. Dal punto di vista della ridondanza codificatoria, questa distribuzione potrebbe essere più o meno efficiente: se le cellule per colore-e-forma replicano l'informazione già portata separatamente dalle cellule per il colore e dalle cellule per la forma, si ha ridondanza nel senso di Shannon. Ma la ridondanza rappresentazionale che qui interessa non riguarda il rapporto tra diverse rappresentazioni della *stessa proprietà* all'interno di un unico livello di analisi; riguarda il rapporto tra rappresentazioni di *livelli diversi* — la forma, il colore, la categoria, il contesto episodico — che concernono tutte, in qualche senso, il medesimo oggetto. Questa seconda forma di ridondanza non può essere valutata come più o meno efficiente senza prima stabilire quale sia il contenuto rispetto al quale l'efficienza viene misurata, il che è precisamente la questione in esame.

All'interno della ridondanza rappresentazionale è opportuno introdurre una distinzione tassonomica che l'analisi successiva utilizzerà in modo sistematico. Si possono distinguere almeno quattro varietà principali, ciascuna delle quali pone problemi filosofici parzialmente distinti e richiede strumenti analitici in parte diversi.

La prima varietà è la ridondanza gerarchica, o compositiva. La stessa proprietà del mondo è rappresentata a livelli diversi della gerarchia cognitiva, con gradi crescenti di astrazione e integrazione. Un orientamento visivo è rappresentato nelle cellule semplici di V1, come gradiente di luminanza orientato in una direzione specifica; il medesimo orientamento partecipa alla rappresentazione di un bordo dell'oggetto a livello di V2 e V4, dove la risposta è invariante rispetto a piccole variazioni di posizione retinica; il medesimo bordo contribuisce alla rappresentazione della forma globale dell'oggetto a livello del cortex inferotemporale, dove l'invarianza si estende a variazioni di scala e rotazione; la medesima forma contribuisce alla rappresentazione della categoria dell'oggetto a livello prefrontale, dove l'invarianza riguarda l'intero spazio di variabilità all'interno di una categoria. Si tratta di una ridondanza di tipo compositivo: le rappresentazioni a livelli superiori contengono, in forma trasformata, l'informazione presente ai livelli inferiori, ma tale trasformazione non è una semplice copia — comporta una perdita dell'informazione idiosincratICA che rende unique quel particolare orientamento, quel particolare bordo, quella particolare forma — e un'aggiunta dell'invarianza che caratterizza il livello superiore.

La seconda varietà è la ridondanza modale-trasversale. Lo stesso oggetto è rappresentato simultaneamente in modalità sensoriali diverse: la forma di una tazza è rappresentata sia visivamente, come pattern di risposta nelle aree del riconoscimento degli oggetti, sia tattilmente, come pattern di risposta nelle aree somatosensoriali, sia in termini di schema motorio di prensione, come configurazione di attivazione nelle aree motorie e premotorie. Queste rappresentazioni non sono compositive nello stesso senso della ridondanza gerarchica: le informazioni provenienti da modalità diverse non si sommano linearmente ma si integrano attraverso meccanismi di congruenza multimodale computazionalmente sofisticati, che si trovano in aree di associazione come il solco temporale superiore e il lobulo parietale inferiore. La convergenza multimodale non produce una rappresentazione amodale che «sussume» quelle modali: produce una struttura di vincoli reciproci tra le rappresentazioni modali che ne aumenta la robustezza e ne stabilizza il contenuto.

La terza varietà è la ridondanza epistemico-prospettica, o aspettuale. Il medesimo fatto del mondo è codificato sia come conoscenza proposizionale esplicita sia come aspettativa implicita procedurale sia come schema percettivo automatizzato. Il fatto che i gatti abbiano quattro zampe è rappresentato come proposizione accessibile alla verbalizzazione esplicita, come aspettativa implicita che si attiva automaticamente alla visione di un gatto di spalle — in attesa che,

girandosi, riveli quattro zampe — e come schema percettivo procedurale che guida il riconoscimento del pattern locomotore felino senza richiedere il recupero consapevole della regola. Queste tre rappresentazioni del medesimo fatto sono accessibili attraverso sistemi cognitivi parzialmente distinti, aggiornabili mediante meccanismi diversi, e vulnerabili a lesioni in aree neurali non coincidenti. Il caso patologico più istruttivo è quello di pazienti con sindrome amnesica che perdono la capacità di formare nuove memorie esplicite ma continuano ad aggiornare le loro abilità procedurali: dimostrano così che la rappresentazione procedurale e quella dichiarativa del medesimo fatto sono sufficientemente indipendenti da potersi dissociare.

La quarta e filosoficamente più sottile varietà è la ridondanza temporale. La stessa struttura cognitiva è mantenuta simultaneamente in diverse scale temporali di rappresentazione: nella memoria di lavoro, come stato attivo nel momento corrente; nella memoria episodica recente, come traccia dell'utilizzo recente del concetto; nella memoria semantica consolidata, come struttura astratta relativamente stabile; negli schemi percettivi e motori, come disposizione automatizzata. Queste non sono semplicemente copie della stessa struttura su diversi supporti: sono versioni con diverse proprietà di accesso, di aggiornabilità, di robustezza e di contestualizzazione. La ridondanza temporale è quella che pone il problema filosofico più diretto riguardo all'identità dei concetti nel tempo, che affronteremo in modo esteso nel Capitolo II.

Questa tassonomia non è esaustiva — vi sono forme di ridondanza che attraversano più dimensioni e non si lasciano facilmente classificare in una categoria sola — ma è sufficiente per articolare le domande principali che la monografia si propone di rispondere.

I.3 — La presupposizione di unicità e le sue conseguenze filosofiche

Per capire perché la ridondanza rappresentazionale costituisce un problema genuino per la filosofia del contenuto, è necessario rendere esplicita la presupposizione che la rende problematica. Chiamo questa presupposizione il principio di unicità del portatore: l'assunzione che per qualsiasi stato mentale con contenuto determinato, vi sia una singola struttura — fisica o funzionale — che costituisce il portatore (*vehicle*) di quel contenuto. Questa presupposizione è così comune nel dibattito che difficilmente emerge come tale; appare piuttosto come un vincolo metodologico minimale, la condizione di possibilità di qualsiasi trattazione precisa del rapporto tra portatori e contenuti.

Il principio di unicità del portatore compare, in forme diverse, nella quasi totalità delle teorie del contenuto influenti. Nella teoria rappresentazionale della mente di Fodor, il contenuto è portato da simboli nel linguaggio del pensiero: ciascun simbolo ha un contenuto

determinato, e la semantica dello stato mentale complesso è funzione compositiva dei contenuti dei simboli che lo compongono. La struttura portante del contenuto è discreta e localizzata: è un token del tipo simbolico appropriato. Nella teoria teologica-informazionale di Dretske, il contenuto di uno stato segnalante è determinato dalla proprietà che lo stato ha la funzione biologica di indicare: c'è una relazione diadica tra lo stato-portatore e la proprietà-contenuto che costituisce il contenuto. In entrambi i casi, il contenuto è individuato da una relazione che coinvolge un singolo portatore.

Le teorie predittive del tipo di Friston e Clark non sfuggono a questa presupposizione, nonostante la loro enfasi sulla distribuzione computazionale. In queste teorie, il contenuto di una rappresentazione gerarchica è portato dalla distribuzione a posteriori — il modello del mondo che il sistema ha costruito — che è calcolata come risultato dell'inferenza bayesiana. Questa distribuzione può essere matematicamente complessa e computazionalmente distribuita attraverso molte aree cerebrali, ma è concettualmente unitaria: è la distribuzione risultante dall'integrazione di tutte le evidenze disponibili, e il suo contenuto è determinato da quella distribuzione nella sua interezza. La molteplicità delle strutture neurali che implementano la computazione è considerata un dettaglio implementativo, non una caratteristica strutturale del contenuto.

Il principio di unicità ha una giustificazione teoretica che vale la pena esplicitare. Se il contenuto è una proprietà relazionale — se, cioè, uno stato ha un certo contenuto in virtù della sua relazione con il mondo — allora sembra naturale individuare il termine interno della relazione come il portatore del contenuto: la relazione corre tra il portatore e il mondo, e il contenuto è il modo in cui il mondo è rappresentato nel portatore. Se invece il portatore è multiplo — se ci sono molti stati ciascuno dei quali ha qualche relazione rilevante con il mondo — allora la relazione di contenuto non è più una relazione diadica tra un portatore e un contenuto, ma una relazione più complessa tra una *struttura* di portatori e il mondo. Questa complessità aggiuntiva richiede una teoria più ricca, e la domanda metodologica è se questa ricchezza aggiuntiva sia giustificata dai fenomeni o sia invece un'inutile complicazione.

La tesi di questa monografia è che sia non solo giustificata ma necessaria, e che il principio di unicità abbia un costo teorico nascosto che emerge non appena si esaminano i fenomeni neurobiologici con sufficiente attenzione. Il costo è il seguente: per mantenere il principio di unicità di fronte all'evidenza neurobiologica della distribuzione, le teorie del contenuto devono o ben identificare il «vero» portatore tra i molti candidati neurali — e abbiamo visto che questo è empiricamente insostenibile — oppure dichiarare che le strutture neurali sono irrilevanti per il contenuto e che la teoria del contenuto opera a un livello di astrazione che è indipendente dai dettagli neurobiologici. Quest'ultima mossa è legittima ma costosa: implica che la teoria del contenuto non

può essere usata per fare previsioni sui correlati neurali del contenuto, il che la priva di una delle sue più importanti risorse esplicative e la rende, in una certa misura, empiricamente vuota.

I.4 — Tre argomenti per la ridondanza costitutiva

Conviene anticipare, in forma sommaria, i tre principali argomenti a sostegno della tesi della ridondanza costitutiva che saranno sviluppati in dettaglio nei capitoli successivi. Presentarli ora in forma compatta permette al lettore di orientarsi nell'architettura dell'intero volume e di valutare l'interdipendenza degli argomenti.

Il primo argomento è quello della necessità funzionale della ridondanza. Si argomenterà, attraverso un'analisi formale della teoria dell'informazione parziale (Capitolo IV), che un sistema cognitivo in grado di rappresentare contenuti stabili, generalizzabili e robusti rispetto al rumore deve necessariamente implementare una struttura di ridondanza costitutiva. L'argomento è di tipo normativo-informativo: non si tratta semplicemente di osservare che i sistemi cognitivi reali sono ridondanti, ma di mostrare che qualsiasi sistema che abbia le proprietà funzionali che caratterizzano la cognizione deve essere strutturalmente ridondante nel senso qui definito. Questa necessità funzionale fornisce una ragione per cui la ridondanza non è un accidente evolutivo ma una caratteristica strutturale di qualsiasi sistema cognitivo sufficientemente ricco.

Il secondo argomento è quello della spiegazione delle dissociazioni. Si argomenterà (Capitolo III) che i dati neuropsicologici sulle dissociazioni tra contenuto semantico integro e contenuto percettivo compromesso, tra identità riconoscibile e valenza emotiva disturbata, tra conoscenza proposizionale e abilità procedurale, sono più coerentemente spiegati dall'ipotesi della ridondanza costitutiva che da qualsiasi alternativa che mantenga il principio di unicità del portatore. Le dissociazioni non sono anomalie da spiegare come casi di perdita parziale di un'unica struttura; sono fenomeni che rivelano la struttura interna della ridondanza stessa, mostrando quali componenti della firma informazionale sono dipendenti da quali strutture neurali.

Il terzo argomento è quello della convergenza descrittiva. Si mostrerà (Capitoli VIII e IX) che la tesi della ridondanza costitutiva, sviluppata attraverso gli strumenti della teoria dell'informazione parziale e della topologia algebrica, converge con descrizioni del contenuto sviluppate attraverso tradizioni filosofiche completamente diverse: la fenomenologia husserliana dell'oggetto intenzionale, la teoria wittgensteiniana del significato come uso, l'approccio enattivista di Maturana e Varela, e la tradizione dell'esternalismo attivo di Clark e Chalmers. Questa convergenza da percorsi indipendenti non è una dimostrazione, ma

costituisce un forte indizio che la struttura rilevata dalla tesi della ridondanza costitutiva corrisponde a qualcosa di genuinamente presente nel fenomeno del contenuto mentale.

I.5 — Delimitazione del campo: cosa questa monografia non sostiene

Per evitare fraintendimenti che potrebbero compromettere la ricezione dell'argomentazione, è opportuno essere espliciti su ciò che la tesi della ridondanza costitutiva non sostiene. Tre interpretazioni erranee sono particolarmente facili da incorrere e vale la pena precluderle fin dall'inizio.

La prima interpretazione erranea è che la ridondanza costitutiva sia una versione del connessionismo o del distributed representation claim nel senso di Rumelhart e McClelland. Le teorie connessioniste sostengono che le rappresentazioni mentali non sono simboli localizzati ma pattern di attivazione distribuiti su reti neurali. Questa tesi riguarda il modo in cui le rappresentazioni sono implementate fisicamente, e la sua versione più forte — supercognitivism — è talvolta interpretata come il rifiuto dell'esistenza di singole rappresentazioni localizzate. Ma la tesi della ridondanza costitutiva non si occupa primariamente dell'implementazione fisica delle rappresentazioni: si occupa della struttura del contenuto. Anche se tutte le rappresentazioni fossero localizzate — anche se il modello della memoria semantica di Fodor, con simboli discreti e localizzati, fosse empiricamente corretto — il problema della ridondanza costitutiva sussisterebbe ancora, perché riguarda la relazione tra rappresentazioni locali multiple, non la struttura interna di ciascuna di esse.

La seconda interpretazione erranea è che la ridondanza costitutiva implichi il rifiuto del realismo semantico o una qualche forma di relativismo del contenuto. Si potrebbe pensare che, se il contenuto è una proprietà relazionale tra rappresentazioni multiple piuttosto che una proprietà assoluta di singole rappresentazioni, allora il contenuto è in qualche senso relativo al sistema cognitivo che lo produce — non vi è un contenuto «oggettivo» che il sistema più o meno accuratamente rappresenta. Ma questo non segue: la struttura di coerenza tra rappresentazioni multiple può essere più o meno fedele a una struttura oggettiva nel mondo, e la distanza tra la firma informazionale di un sistema e la firma di riferimento — che sarà formalizzata nel Capitolo V — misura precisamente questo grado di fedeltà. La ridondanza costitutiva è compatibile con il realismo semantico e, anzi, offre una caratterizzazione più precisa di cosa significhi che un sistema rappresenta accuratamente il mondo.

La terza interpretazione erranea è che la tesi sia riducibile a un olismo del contenuto nel senso di Quine o di Davidson. L'olismo del contenuto

sostiene che il contenuto di qualsiasi credenza o stato mentale dipende dal suo posto nell'intera rete delle credenze del soggetto: il significato di «acqua» per me dipende da tutto ciò che credo riguardo all'acqua, alla chimica, alla sete, e così via. Questa è una tesi sulla dipendenza del contenuto dall'intero sistema di credenze. La tesi della ridondanza costitutiva è più specifica: sostiene che il contenuto di uno stato dipende dalla struttura di coerenza tra le rappresentazioni ridondanti di quello specifico stato — non dall'intero sistema di credenze. Non è un olismo globale ma una tesi sulla struttura locale del contenuto. La relazione tra la tesi della ridondanza costitutiva e l'olismo del contenuto è una questione aperta su cui si tornerà nel Capitolo IX; per ora basti notare che le due tesi sono logicamente indipendenti.

I.6 — Strumenti formali: una mappa preliminare

L'argomentazione di questa monografia fa uso di strumenti formali che provengono da tradizioni diverse e che non sono comunemente familiari ai filosofi del contenuto mentale. È opportuno fornire una mappa preliminare di questi strumenti, indicando per ciascuno il problema specifico che è chiamato a risolvere e il capitolo in cui sarà sviluppato. Una trattazione tecnica completa sarà fornita nei capitoli pertinenti; qui si mira soltanto a orientare il lettore.

Il primo strumento è la teoria dell'informazione parziale (Partial Information Decomposition, PID), sviluppata da Williams e Beer nel 2010 e successivamente raffinata da Griffith, Bertschinger e altri. La PID fornisce un modo formale per decomporre l'informazione mutua tra un insieme di sorgenti $S = \{X_1, \dots, X_n\}$ e un obiettivo Y nelle sue componenti di ridondanza, unicità e sinergia. La ridondanza $\mathcal{R}(\{X_i\}; Y)$ misura l'informazione che più sorgenti forniscono congiuntamente su Y , al di là di quella che ciascuna fornisce individualmente; l'informazione unica $U_i(Y)$ misura l'informazione che solo la sorgente X_i fornisce su Y ; la sinergia $\mathcal{S}(\{1, \dots, n\}; Y)$ misura l'informazione che emerge solo dall'insieme delle sorgenti considerato congiuntamente, che non è fornita da nessuna sorgente individualmente. Questo strumento sarà usato nel Capitolo IV per caratterizzare formalmente la struttura della ridondanza costitutiva come reticolo di atomi informativi.

Il secondo strumento è la geometria delle metriche riemanniane su spazi di distribuzioni di probabilità, nota come geometria dell'informazione nel senso di Amari. Questa geometria fornisce un modo per misurare le distanze tra distribuzioni di probabilità che è sensibile alla struttura dell'informazione, non soltanto alla distanza euclidea tra i valori numerici delle probabilità. Sarà usata nel Capitolo V per definire la firma informazionale di un sistema di rappresentazioni ridondanti e per misurare la distanza tra firme diverse come indicatore del grado di allineamento tra sistemi cognitivi diversi.

Il terzo strumento è la topologia persistente (persistent homology), un metodo della topologia algebrica per estrarre caratteristiche strutturali stabili da dati ad alta dimensionalità. La topologia persistente permette di identificare «loop» e «cavità» nella struttura geometrica di un insieme di dati che persistono attraverso diverse scale di analisi, e che possono essere interpretati come strutture di dipendenza di ordine superiore non riducibili alle dipendenze di ordine inferiore. Sarà usata nel Capitolo V per caratterizzare la struttura topologica della firma informazionale e per distinguere tra firme che hanno la stessa «forma» locale ma strutture globali diverse.

Il quarto strumento è la teoria dei reticoli, o teoria degli ordini parziali, che fornisce il linguaggio naturale per descrivere le gerarchie di anti-chain su cui è costruita la decomposizione PID. I reticoli permettono di formalizzare le relazioni di «più ridondante di» e «più sinergico di» tra diverse strutture di rappresentazioni ridondanti, e di caratterizzare la continuità delle traiettorie nello spazio dei contenuti come traiettorie nello spazio dei reticoli dotato di una metrica appropriata.

Questi quattro strumenti convergono verso una rappresentazione del contenuto mentale come struttura geometrico-informazionale — un oggetto che vive in uno spazio di firme dotato di metrica e topologia — che permette di formulare in modo preciso le domande sulla identità, la variazione e la distanza tra contenuti che le teorie del contenuto tradizionali possono solo descrivere qualitativamente.

I.7 — Struttura del volume

La monografia è organizzata in nove capitoli principali, più un'appendice matematica che raccoglie le definizioni formali principali. I capitoli procedono da un'analisi del problema (Capitoli I–III) attraverso la costruzione del quadro formale (Capitoli IV–V) verso l'applicazione del quadro a problemi filosofici specifici (Capitoli VI–IX).

Il Capitolo II — che segue immediatamente — analizza in dettaglio le quattro varietà di ridondanza introdotte nel presente capitolo, con particolare attenzione alla ridondanza epistemico-prospettica e temporale che pongono i problemi filosofici più sottili riguardo all'identità e alla persistenza del contenuto. Il Capitolo III presenta l'evidenza neuropsicologica delle dissociazioni come banco di prova empirico per le diverse teorie del contenuto. I Capitoli IV e V sviluppano il quadro formale della teoria della ridondanza costitutiva. I Capitoli VI, VII e VIII applicano il quadro al problema dell'errore rappresentazionale, al problema del cambiamento concettuale, e al problema dell'allineamento dei sistemi artificiali. Il Capitolo IX raccoglie le conclusioni filosofiche generali e ne discute le implicazioni per il dibattito sul contenuto mentale.

Capitolo II — Le quattro varietà della ridondanza: analisi e approfondimento

II.1 — Ridondanza gerarchica: la stratificazione del contenuto visivo

La ridondanza gerarchica è quella che la neuroscienza visiva ha caratterizzato con maggiore precisione, e per questa ragione conviene cominciare da essa come caso paradigmatico. Il sistema visivo dei primati offre il caso di studio più studiato di architettura gerarchica nel cervello: una sequenza di aree — dalla retina alle cellule gangliari, al nucleo genicolato laterale del talamo, alla corteccia visiva primaria V1, poi V2, V4, V8, e infine le aree del cortex inferiotemporale come l'area IT posteriore e il giro fusiforme anteriore — in cui ciascuna area riceve l'input dall'area precedente, elabora ulteriori trasformazioni, e trasmette la propria risposta all'area successiva.

Questa architettura gerarchica produce rappresentazioni ridondanti in un senso preciso. Una proprietà visiva semplice come l'orientamento di un bordo — diciamo, un bordo verticale a 90° dalla verticale con una certa posizione retinica (x, y) e una certa frequenza spaziale f — è rappresentata in V1 come pattern di risposta di un insieme di cellule semplici orientate a 90° . La stessa informazione di orientamento è presente, in forma trasformata, nelle risposte delle cellule di V2, che sono selettive per lo stesso orientamento ma mostrano una maggiore invarianza per la posizione retinica. La stessa informazione è ancora presente, più trasformata, nelle risposte dell'area V4, dove la selettività per l'orientamento si combina con la selettività per la curvatura locale e la figura-sfondo. E così via, lungo la gerarchia.

La domanda filosoficamente rilevante è: in che senso tutte queste rappresentazioni sono rappresentazioni «dello stesso orientamento»? La risposta ovvia è che vi è una relazione di dipendenza causale ascendente: le risposte di V2 sono causalmente dipendenti dalle risposte di V1, che a loro volta dipendono causalmente dallo stimolo retinico. Ma questa dipendenza causale non è sufficiente a stabilire l'identità del contenuto. Si consideri che le risposte di V2 dipendono anche dall'attività top-down proveniente dalle aree superiori, dai meccanismi di attenzione che modulano la risposta, dal contesto figurativo in cui il bordo appare. La risposta di V2 a un bordo verticale in un contesto privo di contrasto è sistematicamente diversa dalla risposta allo stesso bordo in un contesto ricco di contrasto; eppure entrambe le risposte sarebbero considerate rappresentazioni «dello stesso bordo verticale». Questo significa che il contenuto della rappresentazione in V2 non è determinato univocamente dalla relazione causale con lo stimolo; è determinato anche dal contesto, dall'attenzione, dalle aspettative.

Formalizzando: sia φ_k il vettore di attivazione dell'area visiva al livello k della gerarchia, e sia s il vettore dello stimolo retinico. Le relazioni tra livelli adiacenti possono essere scritte come:

$$\varphi_{k+1} = f_{k+1}(\varphi_k, c_k, a_k)$$

dove c_k è il vettore del contesto (incluso il feedback top-down dalle aree superiori) e a_k è il vettore dell'attenzione. La funzione f_{k+1} è non lineare e contiene interazioni tra le tre sorgenti di input. Il punto cruciale è che questa equazione mostra che φ_{k+1} non è una funzione semplice di φ_k : è una funzione dell'intera struttura di input $\{\varphi_k, c_k, a_k\}$. Ciò significa che la ridondanza tra φ_k e φ_{k+1} — la quantità di informazione che condividono sullo stimolo s — non è identità ma sovrapposizione parziale con componenti idiosincratice a ciascun livello.

La teoria dell'informazione parziale consente di caratterizzare questa situazione con precisione. Si consideri l'informazione mutua tra la coppia di aree $\{\varphi_k, \varphi_{k+1}\}$ e lo stimolo s . Questa informazione mutua può essere decomposta in:

$$I(\{\varphi_k, \varphi_{k+1}\}; s) = \mathcal{R}(\{k, \{k+1\}; s) + U_k(s) + U_{k+1}(s) + \text{Syn}(\{k, k+1\}; s)$$

dove $\mathcal{R}$ è la ridondanza — l'informazione che entrambe le aree forniscono sullo stimolo — U_k e U_{k+1} sono le informazioni uniche di ciascuna area — l'informazione che un'area fornisce sullo stimolo che l'altra non fornisce — e Syn è la sinergia — l'informazione sullo stimolo che emerge solo dall'analisi congiunta delle due aree.

Questo formalismo permette di distinguere tra diversi regimi della ridondanza gerarchica. In un regime di pura ridondanza ($\text{Syn} \approx 0$, $U_k \approx 0$, $U_{k+1} \approx 0$), l'area superiore non fa che replicare l'informazione dell'area inferiore: il contenuto è identico a tutti i livelli e la gerarchia è ridondante in senso shannoniano. In un regime di pura trasformazione ($\mathcal{R} \approx 0$, $U_k \approx 0$, $U_{k+1} \approx 0$, $\text{Syn} > 0$), ogni livello porta informazione sullo stimolo che non è presente separatamente nei livelli precedenti: il contenuto cambia radicalmente da livello a livello. In un regime misto — quello che l'evidenza neurobiologica suggerisce per il sistema visivo — vi sono sia componenti di ridondanza che componenti di trasformazione, e la firma informazionale della gerarchia ha una struttura di atomi ricca e complessa.

La struttura di atomi della decomposizione PID per una catena gerarchica di n livelli è un oggetto matematico di notevole ricchezza. Per una catena a tre livelli $\{\varphi_1, \varphi_2, \varphi_3\}$ e uno stimolo s , il numero di atomi distinti è dato dal numero di anti-chain del reticolo delle partizioni di $\{1, 2, 3\}$, che è pari a 18 atomi non banali. Ciascuno di questi atomi corrisponde a una componente dell'informazione mutua $I(\{\varphi_1, \varphi_2, \varphi_3\}; s)$ attribuibile a una specifica struttura di ridondanza/sinergia tra le

tre aree. La firma informazionale della gerarchia visiva è dunque un vettore di 18 componenti (per la sola decomposizione a tre livelli; il numero cresce rapidamente con n), e il contenuto del sistema gerarchico nel suo insieme non si riduce a nessuna delle rappresentazioni dei livelli individuali.

II.2 — Ridondanza modale-trasversale: l'integrazione come costituzione del contenuto

La ridondanza modale-trasversale porta al centro della scena un problema filosofico che la ridondanza gerarchica lasciava in ombra: il problema della costituzione del contenuto attraverso l'integrazione di informazioni eterogenee. Nella ridondanza gerarchica, i diversi livelli elaborano informazioni della stessa modalità (visiva, nel caso paradigmatico), semplicemente a gradi crescenti di astrazione. Le trasformazioni che collegano i livelli, per quanto non banali, mantengono una certa continuità del tipo di informazione elaborata. Nella ridondanza modale-trasversale, invece, le diverse rappresentazioni appartengono a modalità qualitativa diversa: la forma visiva di una tazza e la forma tattile della stessa tazza sono codificate in aree anatomicamente distinte, attraverso trasduttori fisicamente diversi, con proprietà di risoluzione spaziale e temporale radicalmente dissimili.

Eppure il contenuto della rappresentazione integrata — «questa tazza, con questa forma, in questa posizione» — sembra essere qualcosa di unitario che trascende le singole modalità. Il fatto che lo stesso oggetto possa essere riconosciuto visivamente e tattilmente, e che il riconoscimento visivo faciliti le aspettative tattili e vice versa, suggerisce che le rappresentazioni modali diverse condividono qualcosa che va oltre la mera correlazione causale con il medesimo oggetto fisico. Questo «qualcosa» è la struttura di coerenza che il quadro della ridondanza costitutiva tematizza formalmente.

La neurobiologia dell'integrazione multisensoriale ha identificato le aree corticali in cui avviene questo processo di integrazione: il solco temporale superiore (STS), che risponde in modo superadditivo alle stimolazioni audiovisive congruenti; il lobulo parietale inferiore, in particolare le aree 40 e 39 di Brodmann, che integra informazioni visuospatiali e somatosensoriali; le aree di associazione frontali, che integrano informazioni da tutte le modalità per guidare l'azione. La risposta superadditiva — il fatto che la risposta alla stimolazione bimodale congruente sia maggiore della somma delle risposte alle stimolazioni unimodali — è il segno neurobiologico che l'integrazione non è semplicemente la somma delle informazioni modali separate, ma produce qualcosa di qualitativamente diverso.

Formalmente, la superaddittività può essere espressa nei termini della teoria dell'informazione parziale come la presenza di sinergia positiva tra le rappresentazioni modali:

$$\text{Syn}(\{\text{vis}, \text{tatt}\}; O) > 0$$

dove O è l'oggetto nel mondo e $\{\text{vis}, \text{tatt}\}$ sono le rappresentazioni visiva e tattile dell'oggetto. La sinergia positiva significa che la conoscenza congiunta della rappresentazione visiva e di quella tattile fornisce informazione sull'oggetto che non è disponibile dall'analisi separata di ciascuna modalità. Questo è precisamente il senso formale in cui l'integrazione multimodale produce qualcosa di qualitativamente nuovo: non una semplice somma ma una struttura emergente.

L'effetto di McGurk è il caso paradigmatico di questa sinergia tra modalità. Quando uno stimolo uditivo /ba/ è presentato simultaneamente a uno stimolo visivo del volto che articola /ga/, il percipiente sente /da/ — una sillaba che non è né quella uditiva né quella visiva, ma una fusione che emerge solo dall'integrazione delle due. Questo fenomeno dimostra che il contenuto del percetto integrato — /da/ — non si trova né nella rappresentazione uditiva né in quella visiva prese separatamente, ma nella struttura della loro relazione reciproca. È un caso di ridondanza costitutiva nel senso più forte: il contenuto è letteralmente prodotto dall'interazione tra le rappresentazioni, non preesiste a essa.

Un altro caso istruttivo è quello della percezione dello spazio corporeo attraverso la visione e la propriocezione. Quando osservo la mia mano muoversi nello spazio visivo, la rappresentazione visiva della posizione della mano e la rappresentazione propriocettiva della posizione dell'arto sono due fonti di informazione sulla medesima quantità fisicamente — la posizione dell'estremità distale del braccio nel sistema di riferimento dello spazio peripersonale. In un soggetto normale, queste due fonti sono altamente congruenti e si rafforzano reciprocamente; in un paziente con lesioni propriocettive, la congruenza è disturbata e il soggetto può vedere la propria mano in una posizione diversa da quella che «sente». La perdita di coerenza tra le rappresentazioni modali — la perdita della struttura di ridondanza modale-trasversale — produce una perturbazione del contenuto del percetto corporeo che non può essere spiegata come la perdita di una singola fonte di informazione.

II.3 — Ridondanza epistemico-prospettica: i livelli del sapere come livelli del contenuto

La ridondanza epistemico-prospettica — la coesistenza di rappresentazioni della stessa realtà in sistemi cognitivi con diversi gradi di esplicitezza, accessibilità e modalità di aggiornamento — è quella che più

direttamente intercetta i problemi classici della filosofia della conoscenza. La distinzione tra conoscenza dichiarativa ed esplicita da una parte, e conoscenza procedurale e implicita dall'altra, è stata tematizzata in modi diversi da Ryle (che la chiama distinzione tra sapere-che e sapere-come), da Searle (che distingue tra intenzionalità originale e intenzionalità derivata nei sistemi cognitivi), da Evans (che distingue entre *connaissance*, *savant* e *croyances implicites*) e da Squire e Zola-Morgan nel contesto della neurobiologia della memoria.

Il problema filosofico specifico che questa varietà di ridondanza pone è quello della relazione tra i diversi livelli di esplicitezza del sapere. Quando affermo che «i cani hanno quattro zampe», esprimo una credenza proposizionale esplicita; quando mi aspetto implicitamente di vedere quattro zampe in un cane che ho parzialmente occultato, metto in mostra un'aspettativa percettiva implicita; quando riconosco automaticamente il pattern locomotore canino senza potere esplicitar la regola che uso, metto in mostra uno schema procedurale. Questi tre stati sembrano riguardare lo stesso fatto del mondo — che i cani hanno quattro zampe — ma appartengono a sistemi cognitivi diversi, si aggiornano attraverso meccanismi diversi, e possono dissodersi in condizioni patologiche.

La questione è se questi tre stati abbiano lo stesso contenuto o contenuti diversi. La risposta ingenua è che abbiano lo stesso contenuto, perché riguardano lo stesso fatto del mondo. Ma questa risposta diventa difficile da mantenere quando si considerano i casi di dissociazione. Un paziente con lesione del lobo temporale mediale sinistro può avere una credenza proposizionale disturbata riguardo alle proprietà degli animali — confondendo i caratteri tipici di gatti e cani — mentre mantiene intatti gli schemi percettivi e motori che guidano il suo comportamento in presenza di cani. Se affermiamo che tutti e tre gli stati hanno lo stesso contenuto, dobbiamo concludere che il paziente ha allo stesso tempo il contenuto corretto (nei sistemi procedurali) e quello scorretto (nel sistema proposizionale); se affermiamo che hanno contenuti diversi, dobbiamo spiegare in che senso si tratta dello stesso fatto del mondo.

La risposta che offre il quadro della ridondanza costitutiva è che tutti e tre gli stati contribuiscono alla struttura di coerenza che costituisce il contenuto complessivo, ma lo fanno in modi diversi e con «pesi» diversi. La firma informazionale del sistema cognitivo per il fatto «i cani hanno quattro zampe» è una struttura di atomi informativi in cui la componente proposizionale, la componente aspettale e la componente procedurale contribuiscono ciascuna con la propria ridondanza, unicità e sinergia rispetto all'insieme dei fatti del mondo rilevanti. Nella condizione di lesione, la componente proposizionale è perturbata ma le altre componenti mantengono la loro struttura; la firma complessiva è degradata ma non annullata, e il comportamento del paziente riflette questa firma degradata — corretto nelle situazioni in cui le componenti intatte sono sufficienti, errato in quelle che richiedono l'elaborazione proposizionale esplicita.

Questa analisi ha una conseguenza importante: il contenuto di uno stato cognitivo non è determinato da una singola componente del sistema, ma dalla struttura complessiva della firma informazionale. Perturbazioni di componenti diverse producono perturbazioni qualitativamente diverse del contenuto. Una lesione che altera la componente proposizionale produce un tipo diverso di deficit rispetto a una lesione che altera la componente procedurale, non perché le componenti abbiano contenuti completamente diversi, ma perché contribuiscono alla firma complessiva in modi diversi. La neuropsicologia dei deficit selettivi è, da questa prospettiva, una mappa delle componenti della firma informazionale.

II.4 — La struttura formale della ridondanza epistemico-prospettica

Il caso della ridondanza epistemico-prospettica si presta particolarmente bene a una formalizzazione che anticipa il quadro sviluppato nel Capitolo IV. Sia Θ un fatto del mondo — diciamo, il fatto che i cani abbiano quattro zampe. Il sistema cognitivo rappresenta Θ attraverso almeno tre tipi di rappresentazione: φ_{prop} (la rappresentazione proposizionale esplicita), φ_{impl} (l'aspettativa percettiva implicita), φ_{proc} (lo schema procedurale). Definiamo il sistema $\Xi = (\varphi_{\text{prop}}, \varphi_{\text{impl}}, \varphi_{\text{proc}})$ come il vettore delle rappresentazioni ridondanti di Θ .

L'informazione mutua totale tra Ξ e Θ è:

$$I(\Xi; \Theta) = I(\varphi_{\text{prop}}, \varphi_{\text{impl}}, \varphi_{\text{proc}}; \Theta)$$

Questa quantità misura quanta informazione il sistema cognitivo complessivo possiede su Θ . Ma questa misura aggregata non cattura la struttura della ridondanza. Per catturarla, è necessario decomporre $I(\Xi; \Theta)$ nei suoi atomi, come si vedrà in dettaglio nel Capitolo IV. In questa sede basti notare che la decomposizione produce, tra gli altri, i seguenti atomi:

$\mathcal{R}(\Theta) \equiv \mathcal{R}(\{\text{prop}\}, \{\text{impl}\}, \{\text{proc}\}; \Theta)$ — la ridondanza trilaterale: l'informazione su Θ che tutte e tre le rappresentazioni forniscono congiuntamente. Questa è l'informazione che rimane accessibile al sistema indipendentemente da quale componente viene interrogata; è la componente più robusta e più «centrale» della rappresentazione di Θ .

$U_{\text{prop}}(\Theta)$ — l'informazione unica della componente proposizionale: l'informazione su Θ che solo φ_{prop} fornisce, e che le altre due componenti non sono in grado di veicolare. È, intuitivamente, l'informazione che si perde quando un paziente ha una lesione selettiva al sistema proposizionale.

$U_{\text{impl}}(\Theta)$ — l'informazione unica della componente implicita: l'informazione su Θ che solo φ_{impl} fornisce. Include, ad esempio, le

aspettative percettive raffinate attraverso l'esposizione a molti esemplari individuali che non sono codificate nella struttura proposizionale.

$U_proc(\Theta)$ — l'informazione unica della componente procedurale: l'informazione su Θ che si manifesta solo nell'esecuzione abile e automatica di comportamenti rilevanti.

$Syn_ \{prop,impl\}(\Theta)$, $Syn_ \{prop,proc\}(\Theta)$, $Syn_ \{impl,proc\}(\Theta)$ — le sinergie bilaterali tra coppie di componenti.

$Syn_ \{prop,impl,proc\}(\Theta)$ — la sinergia trilaterale: l'informazione su Θ che emerge solo dall'analisi congiunta di tutte e tre le componenti e che non è disponibile da nessun sottoinsieme.

La co-informazione totale tra le tre componenti rispetto a Θ è:

$$TC(\Xi \mid \Theta) = \sum_k \{ H(\varphi_k \mid \Theta) - H(\varphi_{prop}, \varphi_{impl}, \varphi_{proc} \mid \Theta) \}$$

Questa quantità è zero se e solo se le tre componenti sono condizionalmente indipendenti dato Θ — il che corrisponderebbe a un sistema in cui i diversi tipi di rappresentazione non si vincolano reciprocamente una volta noto il fatto del mondo — ed è positiva nella misura in cui le componenti si vincolano reciprocamente, cioè nella misura in cui conoscere il valore di una componente riduce l'incertezza sulle altre anche condizionatamente su Θ . Un valore alto di $TC(\Xi \mid \Theta)$ segnala una struttura di coerenza interna ricca tra le rappresentazioni ridondanti — una struttura in cui le rappresentazioni si «parlano» tra loro, si vincolano reciprocamente e si correggono a vicenda. Questa coerenza interna è precisamente la struttura che permette al sistema di mantenere un contenuto stabile nonostante le perturbazioni delle singole componenti.

II.5 — La ridondanza temporale e il problema della persistenza concettuale

La quarta varietà, quella temporale, pone una sfida ulteriore che le prime tre non affrontano direttamente: il problema di come i concetti persistano nel tempo mantenendo la propria identità attraverso aggiornamenti, modificazioni e riorganizzazioni parziali. Questa è la versione cognitiva di un problema antico della metafisica dell'identità: se tutte le componenti di un oggetto vengono sostituite una alla volta, rimane lo stesso oggetto? La risposta, nel caso dei concetti, è evidentemente complicata dal fatto che i concetti non sono oggetti fisici con composizione discreta, ma strutture funzionali che si implementano in substrati fisici soggetti a continua plasticità.

La ridondanza temporale è il fatto che la stessa struttura concettuale esiste simultaneamente in stati di attivazione con scale temporali molto diverse. In un dato momento, il concetto GATTO è attivo nella memoria

di lavoro come rappresentazione contestualizzata del compito in corso (scala temporale di millisecondi-secondi); è recentemente accessibile nella memoria episodica come traccia degli episodi in cui gatti sono stati incontrati nelle ultime ore o giorni (scala temporale di ore-giorni); è consolidato nella memoria semantica come struttura astratta e relativamente stabile del concetto (scala temporale di anni-decenni); è incorporato nel sistema percettivo come schema di riconoscimento automatico (scala temporale di anni, relativamente resistente alla modifica dopo il periodo critico di sviluppo). Queste quattro istanze temporali non sono copie identiche: ciascuna ha proprietà di accesso, contestualizzazione e aggiornabilità caratteristiche.

La neurobiologia della consolidazione mnestica ha mostrato che queste scale temporali corrispondono a meccanismi molecolari e circuitali distinti. I ricordi a breve termine (ore) dipendono da modificazioni delle sinapsi già esistenti attraverso la fosforilazione reversibile di proteine di membrana e la modulazione dei recettori AMPA; i ricordi a lungo termine (giorni-settimane) richiedono la sintesi di nuove proteine e la formazione di nuove sinapsi, mediata dall'attivazione dei fattori di trascrizione come CREB; la memoria semantica consolidata (anni) sembra dipendere da reti di connettività corteccia-corticale che si rendono progressivamente indipendenti dall'ippocampo attraverso il processo di consolidazione sistemica descritto dal modello di Squire e Alvarez. Queste differenze neurobiologiche non sono irrilevanti per la filosofia del contenuto: implicano che le diverse istanze temporali di un concetto hanno substrati fisici diversi, vulnerabilità a lesioni diverse, e dinamiche di aggiornamento diverse.

Il problema della persistenza concettuale si pone in questi termini: quando il contenuto di un concetto viene aggiornato — quando si apprende qualcosa di nuovo sui gatti che modifica la struttura semantica del concetto GATTO — le diverse istanze temporali si aggiornano in modo asincrono. La struttura in memoria di lavoro si aggiorna immediatamente, incorporando il nuovo apprendimento nella rappresentazione contestuale; la traccia episodica si aggiorna con l'episodio corrente; la struttura semantica si aggiorna lentamente attraverso l'accumulazione di molti episodi, in quello che gli studi computazionali sul *complementary learning systems* chiamano il «processo di integrazione lenta»; lo schema percettivo si aggiorna con grandissima lentezza attraverso l'esposizione ripetuta, e in molti casi non si aggiorna significativamente nel corso della vita adulta.

Durante il periodo transitorio in cui le diverse istanze temporali sono in disaccordo tra loro — in cui la struttura semantica registra ancora le vecchie informazioni mentre la memoria episodica ha già incorporato quelle nuove — in che senso si può affermare che il soggetto possiede un unico concetto GATTO? E più in generale: qual è il criterio di identità per i concetti nel tempo?

La risposta che il quadro della ridondanza costitutiva suggerisce è che l'identità del concetto non richiede sincronizzazione perfetta tra le istanze temporali, ma soltanto una struttura di coerenza sufficientemente robusta da mantenere la continuità funzionale del concetto attraverso il suo aggiornamento. La nozione di «coerenza sufficiente» richiede una caratterizzazione formale che sarà fornita nel Capitolo IV attraverso il concetto di firma informazionale e nel Capitolo VII attraverso il concetto di continuità di traiettoria nello spazio dei contenuti.

Anticipiamo qui soltanto l'intuizione di base. Sia $\Xi(t)$ il vettore delle rappresentazioni ridondanti del concetto GATTO al tempo t , con componenti che corrispondono alle diverse scale temporali introdotte sopra. La dinamica di aggiornamento di $\Xi(t)$ può essere descritta, in prima approssimazione, come un sistema di equazioni differenziali accoppiate:

$$d\Xi/dt = F(\Xi, I(t))$$

dove $I(t)$ è l'input al sistema al tempo t — l'insieme degli stimoli e delle esperienze che aggiornano il concetto — e F è la funzione vettore che descrive la risposta del sistema all'input. Questa dinamica è non lineare e fortemente accoppiata: l'aggiornamento di una componente dipende dallo stato delle altre, perché le diverse istanze temporali si vincolano reciprocamente. In particolare, la componente della memoria di lavoro viene continuamente aggiornata dall'input, ma è anche vincolata dalla struttura della memoria semantica, che esercita un effetto di «pull» verso i valori centrali della distribuzione semantica; la memoria semantica, a sua volta, viene lentamente modificata dall'accumulazione delle tracce episodiche, ma è relativamente resistente agli episodi singoli incongruenti.

Sotto ipotesi plausibili sulla struttura di F — in particolare, sotto l'ipotesi che F sia tale da favorire la coerenza interna tra le componenti — il sistema tende verso attrattori stabili che corrispondono a stati di massima coerenza interna tra le componenti. L'identità del concetto nel tempo corrisponde alla persistenza dell'attrattore: il concetto rimane «do stesso» nella misura in cui la sua struttura di attrattore persiste, anche se le istanze specifiche $\Xi(t)$ si spostano continuamente lungo la traiettoria verso e intorno all'attrattore. Il cambiamento concettuale — che sarà studiato in dettaglio nel Capitolo VII — corrisponde a una transizione di fase nella struttura degli attrattori: un cambiamento qualitativo che porta il sistema da un attrattore a un altro, attraverso un periodo transitorio di elevata instabilità e incoerenza interna.

II.6 — Ridondanza patologica e ridondanza insufficiente: le dissociazioni neuropsicologiche come banco di prova

Una delle strategie metodologiche più efficaci per capire l'architettura funzionale di un sistema è l'analisi delle sue dissociazioni patologiche: i casi in cui il sistema funziona normalmente su alcuni compiti ma è compromesso su altri. Nel contesto della ridondanza rappresentazionale, le dissociazioni neuropsicologiche sono particolarmente illuminanti perché mostrano cosa accade quando alcune delle rappresentazioni ridondanti sono danneggiate mentre le altre rimangono intatte.

Se il contenuto fosse identificato con una singola rappresentazione privilegiata, il danno a quella rappresentazione comporterebbe la perdita totale del contenuto, e il danno a qualunque altra struttura sarebbe irrilevante per il contenuto (anche se potesse avere effetti funzionali per altri motivi). Se invece il contenuto è identificato con la struttura di coerenza tra rappresentazioni multiple, il danno a una singola rappresentazione dovrebbe comportare una degradazione del contenuto proporzionale al contributo di quella rappresentazione alla struttura di coerenza — e il sistema dovrebbe continuare a funzionare, in modo parzialmente degradato, grazie alle rappresentazioni rimanenti.

L'evidenza neuropsicologica è coerente in modo sistematico con la seconda previsione, e in modo altrettanto sistematico inconsistente con la prima. Consideriamo tre casi clinici paradigmatici che saranno analizzati in dettaglio nel Capitolo III.

La prosopagnosia pura — l'incapacità selettiva di riconoscere i volti noti, con conservazione del riconoscimento degli oggetti e delle persone attraverso la voce, la andatura e il contesto — risulta tipicamente da lesioni bilaterali del giro fusiforme anteriore. Se il contenuto del riconoscimento facciale fosse localizzato nel giro fusiforme, ci si aspetterebbe una perdita totale della capacità di riconoscimento; invece, i pazienti prosopagnosici possono spesso identificare le persone attraverso caratteristiche non facciali, riconoscere espressioni emotive dai volti (anche se non identificare chi ne è portatore), e mostrare risposte cutanee galvaniche differenziali ai volti familiari rispetto a quelli non familiari — suggerendo che il sistema di riconoscimento affettivo implicito è parzialmente conservato anche in assenza del riconoscimento esplicito. Questo doppio dissociazione tra riconoscimento esplicito e riconoscimento affettivo implicito è precisamente quello che ci si aspetta se il contenuto del riconoscimento facciale è distribuito tra rappresentazioni multiple che possono dissociarsi.

Il fenomeno di Capgras — il delirio in cui il paziente riconosce il volto di una persona cara ma è convinto che sia un impostore — è, in un certo senso, il complementare della prosopagnosia. Nel delirio di Capgras, il riconoscimento visivo del volto è conservato, ma la risposta affettiva implicita è assente: la persona viene identificata come visivamente identica alla persona cara, ma «sentita» come estranea. La spiegazione di Ellis e Young è che nel delirio di Capgras la via del riconoscimento facciale esplicito è intatta ma la via del riconoscimento affettivo — quella

che normalmente produce la risposta galvanica differenziale ai volti familiari — è lesionata. Dalla prospettiva della ridondanza costitutiva, il delirio di Capgras è una manifestazione dell'incoerenza tra due componenti della firma informazionale del riconoscimento facciale: la componente proposizionale esplicita e la componente affettiva implicita si dissociano, producendo una firma incoerente che il sistema cognitivo di ordine superiore «interpreta» come la presenza di un impostore — l'ipotesi che meglio si adatta alla evidenza incoerente disponibile.

La sindrome amnesica del tipo di Korsakoff — in cui la formazione di nuove memorie episodiche è gravemente compromessa mentre le memorie semantiche e procedurali vecchie sono relativamente conservate — illustra la dissociazione tra la componente temporale episodica e le altre componenti della ridondanza temporale. Claparède descrisse nel 1911 il fenomeno di una paziente amnesica che strinse la mano al medico che le si presentava ogni mattina come un estraneo, ma che un giorno, quando il medico nascose nello handshake uno spillo che punge il palmo, rifiutò la mattina successiva di stringergli la mano senza sapere spiegare perché. Questo esperimento, oggi classico, mostra che la componente procedurale-condizionata dell'apprendimento — l'associazione condizionata tra la visione del medico e il timore di essere punti — può formarsi e persistere anche in assenza della componente episodica esplicita. Dalla prospettiva della ridondanza costitutiva, ciò significa che le componenti temporali diverse della firma informazionale sono sufficientemente indipendenti da potersi dissociare in condizioni lesionali.

II.7 — Le implicazioni metodologiche della tassonomia: come studiare la ridondanza

La tassonomia delle quattro varietà di ridondanza sviluppata in questo capitolo ha implicazioni metodologiche rilevanti per il modo in cui si dovrebbero condurre gli studi empirici sulla struttura delle rappresentazioni mentali. Tradizionalmente, la neuroscienza cognitiva ha studiato le rappresentazioni attraverso paradigmi di imaging (fMRI, EEG, MEG) che misurano l'attivazione di singole aree cerebrali in risposta a stimoli specifici. Questi paradigmi sono ben adatti a caratterizzare le rappresentazioni locali — le «firme neurali» di singole aree — ma sono meno adatti a caratterizzare la struttura delle relazioni tra aree, che è precisamente la struttura che il quadro della ridondanza costitutiva indica come rilevante per il contenuto.

Negli ultimi anni, lo sviluppo di metodi di analisi multivariata — Representational Similarity Analysis (RSA), decoding multivariato, misure di connettività funzionale — ha iniziato a spostare l'attenzione dalla attivazione locale alle strutture di relazione tra rappresentazioni. La RSA,

in particolare, confronta le strutture di dissimilarità tra rappresentazioni neurali e strutture di dissimilarità tra rappresentazioni psicologiche o computazionali, consentendo di testare ipotesi sulla corrispondenza tra geometrie rappresentazionali a diversi livelli di analisi. Questo approccio è, metodologicamente, più vicino allo spirito della ridondanza costitutiva di quanto non lo sia il paradigma dell'imaging localizzazionista.

Ciò che manca, tuttavia, è una misura diretta degli atomi della decomposizione PID nei dati neurali — una misura che richiederebbe la stima congiunta delle distribuzioni di probabilità di risposta di molteplici aree simultaneamente, in funzione degli stimoli. Questo è computazionalmente molto costoso e richiede quantità di dati molto maggiori di quelle tipicamente disponibili negli studi di neuroimaging. Il Capitolo VI discuterà come questa limitazione empirica possa essere affrontata attraverso una combinazione di modelli computazionali e misure proxy delle componenti della firma informazionale.

Capitolo III — Le teorie standard del contenuto di fronte alla molteplicità: un esame critico

III.1 — Il problema della granularità e la presupposizione di unicità

Ogni teoria naturalistica del contenuto mentale condivide, almeno implicitamente, quella che si potrebbe chiamare la presupposizione di unicità: l'assunzione che per ogni stato mentale con contenuto determinato vi sia un fatto unitario — una relazione, una funzione, una struttura causale — che fissa quel contenuto in modo non ambiguo e non dipendente da considerazioni contestuali. Questa presupposizione non viene quasi mai tematizzata come tale, proprio perché appare come la condizione ovvia di intelligibilità di qualsiasi teoria del contenuto. Se non vi è un fatto determinato che fissa il contenuto, la nozione di contenuto sembra dissolversi in una vaga evocazione di somiglianze funzionali prive di determinatezza.

Eppure, come si è mostrato nel capitolo precedente, i dati neurobiologici rendono questa presupposizione visibilmente problematica. Se il contenuto «volto di Maria» è simultaneamente rappresentato nel giro fusiforme — come configurazione di caratteri fisiognomici — nell'amigdala — come profilo di rilevanza emotiva — nell'ippocampo — come nodo di una rete di associazioni biografiche — e nelle regioni prefrontali mediali — come elemento di un modello sociale e predittivo

della persona — e se ciascuna di queste aree contribuisce un aspetto distinto e funzionalmente irriducibile del contenuto complessivo, allora non vi è un unico fatto che fissa il contenuto: vi è una struttura di fatti correlati. La domanda da porre non è più «qual è il contenuto di questo stato?» ma «qual è la struttura delle relazioni tra le componenti del sistema che costituisce il contenuto di questo stato?». Questo spostamento di domanda non è cosmetico o terminologico: cambia radicalmente il tipo di teoria che occorre costruire, i criteri di successo cui una tale teoria deve rispondere, e il tipo di evidenza empirica che è rilevante per valutarla.

Il compito di questo capitolo è esaminare in modo sistematico come ciascuna delle teorie principali del contenuto tenti — implicitamente o esplicitamente — di trattare la molteplicità rappresentazionale, e per quale ragione in ciascun caso il tentativo fallisca o generi difficoltà che la teoria originaria non era attrezzata ad affrontare. L'obiettivo non è confutare globalmente nessuna di queste teorie: tutte catturano aspetti genuini e non trascurabili del fenomeno del contenuto. L'obiettivo è mostrare che la presupposizione di unicità costituisce in ciascuna un punto cieco sistematico, e che le difficoltà che le teorie incontrano nella zona del punto cieco sono intimamente connesse all'impossibilità di trattare adeguatamente la ridondanza costitutiva.

Vale la pena, prima di entrare nell'analisi delle teorie specifiche, articolare più precisamente la struttura del problema. La presupposizione di unicità non è una singola tesi ma un nodo di assunzioni interconnesse. La prima assunzione è quella che potremmo chiamare singolarità del portatore: ogni contenuto mentale determinato è portato da un singolo stato mentale, anche se questo stato può essere fisicamente complesso o distribuito. La seconda assunzione è quella di trasparenza strutturale: la struttura del contenuto rispecchia la struttura del portatore, nel senso che relazioni di contenuto (implicazione, incompatibilità, somiglianza) corrispondono a relazioni strutturali tra portatori. La terza assunzione è quella di indipendenza contestuale: il fatto che fissa il contenuto di uno stato non dipende dal contesto relazionale in cui quello stato si trova rispetto ad altri stati dello stesso sistema.

Queste tre assunzioni formano un sistema coerente che ha radici profonde nella tradizione della filosofia del linguaggio, dove si è sviluppata la nozione di proposizione come oggetto astratto che funge da significato delle frasi e da contenuto delle credenze. La proposizione è singolare (c'è una proposizione, non molte, espressa da una frase con un dato significato), è strutturalmente trasparente (la struttura logica della proposizione è leggibile dalla struttura sintattica della frase), e contestualmente indipendente (la proposizione espressa da una frase non cambia in funzione di quale altra frase sia espressa contemporaneamente). Il trasferimento di questi presupposti dalla semantica formale alle teorie naturalistiche del contenuto mentale è avvenuto in modo largamente implicito, come se la struttura del

linguaggio potesse fungere da modello per la struttura del pensiero senza ulteriori giustificazioni.

Il contributo specifico dei dati neurobiologici discussi nel Capitolo II è proprio quello di mostrare che queste tre assunzioni non sono soddisfatte nel caso della cognizione biologica. La singolarità del portatore è violata perché lo stesso contenuto è portato da stati multipli non sovrapposti. La trasparenza strutturale è violata perché la struttura dei portatori è eterogenea — diversi formati di codifica, diverse dinamiche di aggiornamento, diverse sensibilità contestuali — in modo non rispecchiato dalla struttura unitaria del contenuto attribuito. L'indipendenza contestuale è violata perché la struttura di coerenza tra le rappresentazioni multiple — e quindi il contenuto complessivo — dipende dal contesto cognitivo in cui il sistema si trova.

III.2 — La teleosemantica di Millikan: funzioni proprie e il problema del livello di descrizione

III.2,1 — Il programma teleosemantic nella sua versione canonica

Nel sistema elaborato da Ruth Garrett Millikan nel corso di tre decenni di lavoro — da *Language, Thought, and Other Biological Categories* del 1984 fino agli scritti più recenti sulla comprensione categoriale — il contenuto di una rappresentazione è determinato dalla condizione che essa deve tracciare affinché assolva la propria funzione propria. La funzione propria non è la funzione che un dispositivo di fatto svolge, né la funzione che sarebbe utile che svolgesse: è la funzione che ha conferito al tipo di dispositivo cui appartiene la sua storia selettiva, vale a dire la funzione che spiega causalmente perché quel tipo di dispositivo si è moltiplicato nell'ambiente di selezione rilevante. Un cuore pompa il sangue non perché ci sia qualcosa di intrinsecamente pompante in esso, ma perché i cuori che pompa sangue hanno contribuito alla sopravvivenza degli organismi che li possedevano, e questa storia causale ha selezionato il tipo.

Il meccanismo specifico attraverso cui il contenuto viene determinato si articola in due componenti. Vi è un dispositivo produttore (Producer) che genera rappresentazioni, e un dispositivo consumatore (Consumer) che utilizza quelle rappresentazioni per orientare il comportamento. Il contenuto di una rappresentazione *R* è determinato da ciò che *R* deve tracciare — quale condizione del mondo deve corrispondere — affinché il consumatore di *R* possa assolvere la sua funzione propria. Se il consumatore si comporta come la sua storia selettiva lo ha conformato a

comportarsi quando riceve R, l'organismo otterrà risultati biologicamente adattativi se e solo se la condizione tracciata da R è soddisfatta nel mondo. Questa condizione di adeguamento — la condizione del mondo che permette al consumatore di fare il suo lavoro — è il contenuto della rappresentazione.

La potenza di questa proposta risiede nel fatto che essa incorpora naturalmente la possibilità dell'errore rappresentazionale senza fare appello a relazioni normative primitive o a intenzionalità irriducibile. Una mappa interna può rappresentare erroneamente la posizione di una risorsa perché la condizione che essa deve tracciare — la presenza della risorsa nella posizione indicata — non è soddisfatta, anche se la mappa si è formata causalmente a partire da un certo input. La misrappresentazione non è un'anomalia della causalità: è la condizione in cui il consumatore si comporta come se la condizione fosse soddisfatta, ma essa non lo è.

Millikan distingue due sensi di contenuto corrispondenti ai due poli del meccanismo rappresentazionale. Il contenuto normale del produttore è la proprietà che i produttori di quel tipo devono tracciare affinché producano token che permettono al consumatore di fare il suo lavoro. Il contenuto normale del consumatore è la condizione che deve essere soddisfatta nel mondo perché il consumatore, comportandosi normalmente, produca effetti biologicamente adattativi. Nei sistemi meglio integrati, questi due sensi di contenuto convergono, e si può parlare semplicemente del contenuto della rappresentazione. Ma nei sistemi patologici o nelle condizioni non ordinarie di utilizzo, i due sensi possono divergere, producendo le anomalie semantiche che la teoria deve saper classificare.

III.2,2 — La sfida della molteplicità: cinque difficoltà strutturali

Quando si applica lo schema di Millikan alla ridondanza rappresentazionale, non emerge un singolo problema ma una cascata di difficoltà interconnesse, ciascuna delle quali merita trattazione autonoma perché rivela un aspetto diverso dell'incompatibilità tra la teleosemantica e la struttura distribuita della cognizione biologica.

La prima difficoltà è quella del livello di descrizione, che si è anticipata nella versione precedente di questo capitolo e che ora conviene articolare con maggiore precisione. Il problema non è soltanto che vi sono molti livelli a cui un meccanismo neurale può essere descritto, ciascuno con la propria funzione propria candidata. Il problema è che la storia selettiva rilevante per determinare la funzione propria di un sistema neurale non è, di regola, la storia di quel sistema considerato in isolamento, ma la storia del sistema in relazione agli altri sistemi con cui interagisce. La funzione propria del giro fusiforme non può essere determinata senza fare riferimento alla funzione del sistema di cui fa parte — il sistema di

riconoscimento delle identità personali — e quest'ultima funzione è essa stessa relazionale, dipende cioè dalla struttura delle connessioni tra il giro fusiforme, l'amigdala, l'ippocampo e la corteccia prefrontale.

Consideriamo questo punto mediante un caso controesempio costruito con cura. Immaginiamo un sistema artificiale che incorpori un rilevatore di bordi seguito da un rilevatore di pattern e da un classificatore binario. Il rilevatore di bordi produce rappresentazioni di caratteristiche locali dell'input; il rilevatore di pattern produce rappresentazioni di configurazioni spaziali di caratteristiche; il classificatore produce un'etichetta binaria (volto/non-volto). Se questo sistema viene selezionato in un ambiente di selezione artificiale in cui viene premiato ogni volta che il classificatore etichetta correttamente i volti, qual è la funzione propria di ciascun componente? La risposta immediata è che la funzione del rilevatore di bordi è rilevare bordi, la funzione del rilevatore di pattern è rilevare pattern, e la funzione del classificatore è classificare volti. Ma questa risposta è sbagliata in senso preciso: la funzione del rilevatore di bordi non è rilevare bordi in quanto tali, ma rilevare bordi che contribuiscono al riconoscimento di volti, perché soltanto questa descrizione rende ragione della sua storia selettiva. La funzione propria di ogni componente è internamente relazionale — dipende dal ruolo del componente nel sistema più ampio — e questo rende impossibile determinare la funzione propria di ciascun componente indipendentemente dalla struttura del sistema.

Ora si veda come questo problema si radica nel sistema biologico reale. Il giro fusiforme anteriore risponde in modo preferenziale ai volti, ma non in modo esclusivo: risponde anche ad altri oggetti di riconoscimento fine-grained, come le automobili negli esperici con esperti di auto, e come gli uccelli nei confronti di ornitologi esperti. La «funzione propria» del giro fusiforme non è riconoscere i volti in quanto tali, ma è contribuire al riconoscimento fine-grained degli oggetti di expertise, e questa funzione è determinata — nella storia evolutiva e di sviluppo — dalla struttura delle connessioni neurali che collegano il giro fusiforme alle aree sensoriali di basso livello e alle aree di ordine superiore. Cambiando le connessioni, cambierebbe la funzione propria, e quindi cambierebbe il contenuto delle rappresentazioni che il giro fusiforme produce. Ma le connessioni sono precise quello che le rendono parte di un sistema ridondante. La ridondanza è, in altri termini, costitutiva della funzione propria, non un fenomeno accidentale che la funzione propria descrive.

La seconda difficoltà riguarda la relazione tra ridondanza e degenerazione biologica. Nel vocabolario di Gerald Edelman — che ha sviluppato il concetto di degenerazione come caratteristica strutturale fondamentale dei sistemi neurali selezionati — la degenerazione è la capacità di elementi o strutture strutturalmente diversi di produrre lo stesso output o risultato funzionale. La degenerazione è diversa dalla ridondanza nel senso tecnico ristretto: la ridondanza in quel senso è la

proprietà di componenti identici o quasi identici di produrre lo stesso output, il che è comune nella tecnologia di ingegneria ma raro nei sistemi biologici. La degenerazione biologica consiste invece nell'avere strutture fisicamente diverse — diverse popolazioni neurali, diverse connessioni, diverse dinamiche — che contribuiscono alla medesima funzione. Un sistema biologico è, in senso edelmaniano, degenerato quando il medesimo output funzionale può essere ottenuto attraverso percorsi causali alternativi che non si sovrappongono fisicamente.

Questa distinzione è devastante per la teleosemantica di Millikan. Se il medesimo contenuto può essere portato da strutture fisicamente diverse, e se la funzione propria di una struttura è determinata dalla sua storia selettiva, allora strutture con la stessa storia selettiva di tipo ma con strutture fisiche diverse avranno la medesima funzione propria e lo stesso contenuto. Fin qui nessun problema: questo è esattamente il tipo di realizzazione multipla che il funzionalismo filosofico ha sempre invocato. Ma il problema sorge quando le strutture degenerate non hanno la medesima storia selettiva di tipo, ma producono lo stesso output funzionale attraverso percorsi storici diversi. In tal caso, la teleosemantica di Millikan non garantisce che esse abbiano lo stesso contenuto, eppure la biologia e l'esperienza fenomenologica suggeriscono che producono rappresentazioni dello stesso tipo.

Il caso paradigmatico è il sistema visivo danneggiato e plasticamente riorganizzato. Dopo un danno alla corteccia visiva primaria, alcune funzioni visive possono essere recuperate attraverso percorsi corticali alternativi — il cosiddetto visus cieco o la plasticità corticale post-danno. Le aree che assumono la nuova funzione visiva non hanno la stessa storia selettiva di tipo delle aree danneggiate: si sono sviluppate evolutivamente per svolgere funzioni diverse. Eppure esse producono rappresentazioni che l'organismo utilizza per orientarsi visivamente nel modo in cui usava le rappresentazioni dell'area danneggiata. La teleosemantica di Millikan non ha i mezzi per attribuire lo stesso contenuto alle vecchie e alle nuove rappresentazioni, a meno di non introdurre un concetto di «funzione propria indiretta» che mina la pretesa naturalistica dell'impresa.

La terza difficoltà è quella che si potrebbe chiamare il problema della sincronizzazione funzionale. In un sistema ridondante, le rappresentazioni multiple non producono i loro effetti funzionali in modo indipendente: i loro effetti si sincronizzano in un pattern di attivazione coerente che è esso stesso portatore di informazione. Le oscillazioni gamma nella gamma dei 40 Hz — il meccanismo di binding proposto da Wolf Singer e Francis Crick come soluzione al problema del binding percettivo — sono il caso più studiato, ma non l'unico. Quando due aree neurali che rappresentano aspetti complementari di un oggetto si sincronizzano nelle loro oscillazioni, questa sincronizzazione produce una struttura di coerenza temporale che porta informazione sulla appartenenza comune degli aspetti allo stesso oggetto. La

sincronizzazione stessa è un portatore di informazione, non un semplice meccanismo di trasmissione.

Per la teleosemantica di Millikan, questo crea un problema specifico: qual è la funzione propria della sincronizzazione? La sincronizzazione non è un dispositivo rappresentazionale nel senso ordinario di Millikan: non produce token che vengono consumati da un consumatore. È una relazione dinamica tra dispositivi rappresentazionali distinti, una relazione che porta informazione sull'appartenenza comune. Se si tenta di applicare lo schema millikaniano alla sincronizzazione, si deve identificare un produttore della sincronizzazione e un consumatore della sincronizzazione — ma la sincronizzazione non ha un produttore localizzabile: è il risultato emergente di due sistemi che oscillano in fase. Il consumatore, parimenti, non è un sistema localizzabile che «legge» la sincronizzazione: è la struttura globale del sistema che risponde diversamente alla sincronizzazione e alla desincronizzazione. La teleosemantica, progettata per stati discreti con produttori e consumatori localizzabili, non è strutturalmente adatta a trattare le relazioni dinamiche emergenti come portatori di contenuto.

La quarta difficoltà riguarda l'incoerenza rappresentazionale come stato semanticamente significativo. In un sistema ridondante, è possibile che le diverse rappresentazioni di uno stesso oggetto siano parzialmente incoerenti — che le loro condizioni di soddisfazione si sovrappongano solo parzialmente, o che alcune componenti del sistema aggiornino le loro rappresentazioni più velocemente di altre in risposta a nuove informazioni, producendo un transitorio stato di incoerenza. Questo tipo di incoerenza transitoria è semanticamente significativa: è la struttura cognitiva dell'incertezza, dell'ambiguità, dell'oscillazione percettiva come nelle figure reversibili. La figura di Necker — il cubo che si può vedere in due configurazioni di profondità alternative — è un caso paradigmatico: il sistema visivo mantiene due rappresentazioni alternative, parzialmente incoerenti, dello stesso input, e oscilla tra esse. La struttura dell'oscillazione — la sua periodicità, la sua sensibilità all'attenzione, la sua modulazione da parte delle aspettative — è parte del contenuto dell'esperienza di ambiguità.

La teleosemantica di Millikan può attribuire contenuti alle singole rappresentazioni alternative — ciascuna ha la funzione di tracciare una delle configurazioni di profondità, e questa funzione è determinata dalla storia selettiva del sistema — ma non ha i mezzi per attribuire contenuto alla struttura dell'oscillazione tra esse. L'oscillazione non è la funzione propria di nessun sistema, perché nessun sistema è stato selezionato per oscillare tra alternative incoerenti come stato normale: l'oscillazione è il residuo di un sistema selezionato per risolvere l'ambiguità, che si trova in un ambiente (quello del cubo di Necker) progettato artificialmente per frustrare la risoluzione. Il contenuto dell'ambiguità — il fatto che il sistema sia in uno stato di incertezza tra due alternative — emerge dalla

struttura delle relazioni tra le rappresentazioni in oscillazione, non dalla funzione propria di nessuna di esse individualmente.

La quinta difficoltà è la più fondamentale ed è connessa al problema del livello teleologico appropriato. Millikan distingue la teleologia diretta, quella in cui una struttura serve una funzione nell'organismo attuale, dalla teleologia evolutiva, in cui una struttura è il risultato di una storia selettiva che ha premiato organismi con strutture simili. Ma nella cognizione umana, le rappresentazioni mentali possono avere contenuti che non hanno mai svolto alcuna funzione adattativa nella storia evolutiva della specie. I concetti matematici astratti, le rappresentazioni di situazioni controfattuali, i pensieri su entità non esistenti — tutte queste categorie pongono difficoltà alla teleosemantica che Millikan ha affrontato con la nozione di funzione propria derivata. Ma la ridondanza costitutiva aggiunge una difficoltà ulteriore: anche quando i contenuti sono biologicamente rilevanti, la struttura specifica della ridondanza — il modo particolare in cui le rappresentazioni multiple si combinano per costituire quel contenuto — non ha una spiegazione teleologica diretta. È possibile che la ridondanza sia stata selezionata come meccanismo di robustezza generale, senza che la struttura specifica della ridondanza per ciascun contenuto abbia essa stessa una storia selettiva. In questo caso, la struttura della ridondanza costitutiva — la struttura che, secondo la teoria qui difesa, fissa il contenuto — non è spiegabile in termini teleologici, il che mina la pretesa naturalistica della teleosemantica quando applicata alla teoria della ridondanza.

III.2,3 — Una risposta difensiva: la teleosemantica sistemica e la sua valutazione

Anticipando queste difficoltà, si potrebbe costruire una versione difensiva della teleosemantica che risponda alle obiezioni mediante un allargamento del concetto di funzione propria dal livello dei meccanismi individuali al livello dei sistemi. Nella versione sistemica, la funzione propria di una rappresentazione non è determinata dal ruolo del meccanismo che la produce all'interno del sistema, ma dal ruolo dell'intero sistema nella vita dell'organismo. Il contenuto della rappresentazione del volto di Maria sarebbe determinato dalla funzione propria del sistema cognitivo di riconoscimento dei conspecifici — un sistema che include il giro fusiforme, l'amigdala, l'ippocampo e le regioni prefrontali come componenti interdipendenti — e non dalla funzione propria di ciascun componente individualmente.

Questa versione sistemica ha una certa attrattiva: evita il problema del livello di descrizione assegnando la priorità al livello sistemico, e dissolve il problema delle funzioni indipendenti ammettendo che i componenti hanno funzioni essenzialmente relazionali che si definiscono solo nell'ambito del sistema. Ma la versione sistemica paga un prezzo teorico considerevole che riduce la sua capacità di fare il lavoro per cui è stata costruita.

Il primo problema è che la versione sistemica non offre una risposta al problema della misrappresentazione sistematica. Se il contenuto è determinato dalla funzione propria del sistema, e il sistema funziona come la storia selettiva lo ha conformato a funzionare, allora il sistema rappresenta correttamente quando funziona normalmente. Ma cosa succede quando il sistema produce rappresentazioni che sono sistematicamente ingannevoli — non per malfunzionamento ma per struttura? Il caso paradigmatico è quello delle illusioni ottiche: la percezione di Müller-Lyer, in cui due segmenti di uguale lunghezza sono percepiti come di lunghezza diversa, persiste anche dopo che il soggetto conosce l'illusione. Il sistema visivo funziona normalmente — non è danneggiato — ma produce una rappresentazione che non corrisponde alla struttura del mondo. La teleosemantica sistemica deve dire che il contenuto della rappresentazione di Müller-Lyer è il contenuto corretto dato la funzione del sistema, il che significa che il soggetto che percepisce i segmenti come di lunghezza diversa sta percependo correttamente, e la discrepanza con la realtà non è un errore ma una caratteristica del sistema. Questa conclusione è controintuitiva in modo problematico.

Il secondo problema della versione sistemica è che essa rende il contenuto degli stati individuali dipendente dalla struttura del sistema nel suo insieme, il che mina la granularità esplicativa della teoria. Se il contenuto della rappresentazione prodotta dal giro fusiforme è determinato dalla funzione propria del sistema cognitivo di riconoscimento delle identità personali, allora due sistemi con giri fusiformi identici ma con strutture sistemiche diverse — strutture diverse nell'amigdala, nell'ippocampo, o nelle regioni prefrontali — potrebbero avere giri fusiformi che producono rappresentazioni con contenuti diversi, a parità di struttura fisica e di profilo di risposta del giro fusiforme stesso. Questo risultato implica che il contenuto delle rappresentazioni neurali non è supervenientemente determinato dalla struttura fisica locale, ma dipende dalla struttura globale del sistema — il che è precisamente la tesi della ridondanza costitutiva che si difende in questa monografia. La teleosemantica sistemica, nel tentativo di salvare la propria impresa allargando il livello di analisi, converge verso la posizione ridondantista, riconoscendo implicitamente l'insufficienza del livello locale.

III.2,4 — Il contributo positivo della teleosemantica: funzione evolutiva e struttura della firma

Nonostante le difficoltà esposte, la teleosemantica di Millikan offre un contributo positivo che il quadro della ridondanza costitutiva non può ignorare. Questo contributo consiste nell'idea che la struttura delle rappresentazioni biologiche non sia arbitraria ma rifletta le pressioni selettive che hanno modellato l'architettura cognitiva nel corso dell'evoluzione. Anche se la funzione propria non determina il contenuto nel modo diretto proposto da Millikan, essa determina la struttura della

firma informazionale — la configurazione degli atomi di ridondanza, unicità e sinergia che caratterizza il sistema rappresentazionale per un dato dominio di contenuto.

Si consideri, a titolo di illustrazione, la struttura della ridondanza nella percezione del movimento biologico. Gli esperimenti di Johansson con i punti luce applicati alle articolazioni di un attore in movimento hanno mostrato che il sistema visivo riconosce immediatamente un pattern di movimento biologico da soli dodici punti luce, a prescindere dalla struttura morfologica dettagliata del corpo. Questo riconoscimento coinvolge almeno tre sistemi distinti: il sistema di analisi del movimento optico nell'area MT/V5, il sistema di riconoscimento delle forme biologiche nell'area STS, e il sistema di simulazione motoria nelle regioni premotorie. Ciascuno di questi sistemi contribuisce una componente distinta della rappresentazione del movimento biologico: la componente cinematica (velocità e traiettorie dei punti), la componente configurazionale (la struttura spaziale dell'articolazione), e la componente motoria-simulativa (la rappresentazione interna della possibilità di produrre quel movimento). La struttura di ridondanza tra questi tre sistemi — quanta informazione è condivisa, quanta è unica a ciascuno, quanta emerge solo dalla loro combinazione — riflette le pressioni selettive su un animale sociale che deve riconoscere rapidamente e accuratamente i movimenti dei conspecifici. La teleosemantica, opportunamente integrata, può spiegare perché quella particolare struttura di ridondanza è evolutivamente stabile, anche se non può determinare il contenuto a partire dalla funzione propria.

III.3 — La teoria informazionale di Dretske: causalità, indicazione e il problema della molteplicità

III.3.1 — Il sistema dretsiano nella sua articolazione matura

La teoria dell'indicazione informazionale di Fred Dretske, elaborata con crescente rigore da *Knowledge and the Flow of Information* del 1981 fino a *Naturalizing the Mind* del 1995, parte dal concetto shannoniano di informazione ma lo reinterpreta in chiave semantica attraverso l'introduzione della relazione di indicazione nomologica. Uno stato segnalante *s* indica la proprietà *F* dell'oggetto *x* se e solo se esiste una regolarità nomologica — una legge della natura o una dipendenza causale affidabile — in virtù della quale *s* si produce se e solo se *F* è istanziata in *x*. L'indicazione è dunque una relazione causale-nomologica che garantisce la trasmissione affidabile di informazione dall'oggetto al sistema che lo rappresenta.

Il passo cruciale da indicazione a contenuto rappresentazionale avviene attraverso l'introduzione del concetto di funzione segnalante, che ricorda la funzione propria millikaniana ma si distingue da essa per il meccanismo di acquisizione. Uno stato mentale m ha la funzione segnalante di indicare F se m ha acquisito il ruolo causale nell'organismo di portare l'informazione che F , e questo ruolo causale è stato acquisito attraverso un processo di apprendimento (nella teoria dretsiana del contenuto concettuale) o di selezione (nella teoria dretsiana del contenuto percettivo). Il contenuto m è determinato dalla proprietà F che m ha la funzione segnalante di indicare: m rappresenta F , e la sua condizione di soddisfazione è che F sia istanziata nel generatore rilevante.

La teoria dretsiana ha il merito distintivo di trattare in modo naturale il problema del comportamento intenzionale: un organismo si comporta come se credesse che F perché il suo stato interno m , che ha la funzione segnalante di indicare F , causa il comportamento attraverso le sue proprietà di ruolo causale. La spiegazione del comportamento intenzionale non fa appello a proprietà semantiche primitive: fa appello alla funzione causale di m nell'organismo, e questa funzione causale è ridotta alle relazioni nomologiche tra l'organismo e il suo ambiente.

III.3,2 — Il problema della molteplice indicazione

Il primo punto di attrito tra la teoria dretsiana e la ridondanza rappresentazionale è quello che si potrebbe chiamare il problema della molteplicità delle indicazioni. In un sistema ridondante, molte strutture neurali distinte indicano proprietà diverse dello stesso oggetto, e la domanda «qual è il contenuto del sistema?» non ha una risposta univoca nei termini di una singola relazione di indicazione.

Consideriamo il caso concreto della percezione visiva di un limone. La struttura L_1 nel cortex visivo primario indica la presenza di una macchia gialla di certe dimensioni angolari in una certa posizione del campo visivo. La struttura L_2 nell'area V4 indica la presenza di una superficie di colore giallo saturo con certa riflettanza. La struttura L_3 nell'area IT ventrale indica la presenza di un oggetto della forma caratteristica degli agrumi. La struttura L_4 nell'ippocampo indica la presenza di un oggetto associato al contesto di una cucina italiana o di un mercato. La struttura L_5 nelle regioni orbitofrontali indica la presenza di un oggetto associato a sapori acri e odori caratteristici. Ciascuna di queste strutture indica, in senso nomologico dretsiano, una proprietà diversa del medesimo oggetto fisico.

La domanda «qual è il contenuto della rappresentazione del limone?» riceve da Dretske la risposta che il contenuto è determinato dalla proprietà che lo stato ha la funzione segnalante di indicare. Ma quale delle cinque strutture ha la funzione segnalante rilevante? E se la risposta è «tutte e cinque hanno ciascuna la propria funzione segnalante», allora il

contenuto della rappresentazione del limone non è un contenuto unitario ma una pluralità di contenuti parziali, ciascuno corrispondente a una delle proprietà indicate. Questo è il problema dell'indicazione plurima nella sua forma più netta.

Dretske ha risposto a questo tipo di obiezione sostenendo che il contenuto rilevante è quello della struttura che ha la funzione di indicare la proprietà in virtù della quale il comportamento dell'organismo è prodotto. Quando un organismo afferra un limone per preparare una bevanda, la struttura causalmente rilevante per spiegare quel comportamento è quella che indica le proprietà dell'oggetto rilevanti per l'azione — la sua manipolabilità, il suo aroma, il suo contenuto di succo. Le strutture che indicano proprietà di basso livello come la macchia gialla retinotopica sono causalmente connesse al comportamento attraverso la mediazione delle strutture di ordine superiore, ma non sono «le» strutture il cui contenuto spiega direttamente il comportamento.

Questa risposta è parzialmente soddisfacente ma introduce una nuova difficoltà: la nozione di «struttura causalmente rilevante per il comportamento» non è univoca. Il comportamento di afferrare un limone è causato da tutta la catena di strutture dall'input sensoriale all'output motorio, e la scelta della struttura causalmente «più rilevante» richiede criteri che la teoria dretsiana non fornisce in modo indipendente. Si apre così il problema che si potrebbe chiamare dell'indeterminatezza causalista: la teoria dretsiana non determina univocamente il contenuto perché non determina univocamente quale struttura causale sia quella che porta il contenuto rilevante per la spiegazione del comportamento.

III.3,3 — Il problema della riconfigurazione contestuale

Una seconda difficoltà specifica alla teoria dretsiana emerge considerando la variabilità contestuale della struttura delle rappresentazioni ridondanti. Come si è osservato nel Capitolo II, la struttura di coerenza tra le rappresentazioni multiple di uno stesso oggetto varia in funzione del contesto cognitivo: l'insieme di strutture attivate, il profilo delle loro dipendenze reciproche, e la loro contribuzione relativa al comportamento cambiano in modo sistematico con il compito in atto, le aspettative del soggetto, e lo stato di attenzione del sistema.

Si costruisca un caso controesempio preciso. Un sommelier esperto che degusta un vino attiva, per lo stesso stimolo olfattivo, una struttura di rappresentazioni ridondanti molto diversa da quella attivata da un neofita. Il sommelier attiva L_1 (le caratteristiche fisico-chimiche dell'odore: esteri, terpeni, composti solfurei), L_2 (il profilo varietale: Cabernet Sauvignon, Merlot, Sangiovese), L_3 (il profilo regionale: Bordeaux, Toscana, Napa Valley), L_4 (l'annata e il suo legame con le condizioni climatiche), L_5 (la storia di abbinamento con i cibi), e L_6 (la

valutazione estetica integrata in un contesto di competenza tecnica). Il neofita attiva L_1 e forse L_5 in forma rudimentale, ma non L_2 - L_4 e solo parzialmente L_6 .

Entrambi i soggetti ricevono lo stesso stimolo fisico, e per entrambi le strutture attivate indicano, ciascuna in senso dretsiano, le medesime proprietà fisiche del vino: le stesse molecole odorose, le stesse composizioni chimiche. Eppure il contenuto della rappresentazione è diverso: il sommelier rappresenta il vino come un Château Margaux del 2005, mentre il neofita lo rappresenta come un vino rosso con caratteristiche generiche. La differenza di contenuto non è spiegabile in termini di differenze nelle funzioni segnalanti delle strutture fisicamente presenti, perché le strutture che indicano le proprietà varietali e regionali sono fisicamente assenti nel sistema del neofita. Il contenuto differisce perché la firma informazionale — la struttura di ridondanza, unicità e sinergia tra le rappresentazioni multiple — è radicalmente diversa nei due sistemi. Questo è un caso in cui la teoria dretsiana, che determina il contenuto dalla funzione segnalante di strutture individuali, non può catturare la differenza di contenuto tra due sistemi che ricevono lo stesso input fisico ma lo processano con architetture rappresentazionali diverse.

III.3,4 — Il problema del disjunction e la ridondanza come risposta

Il problema del disjunction — uno dei problemi classici della teoria dretsiana — assume una forma nuova e particolarmente acuta nel contesto della ridondanza rappresentazionale. Il problema del disjunction originale è il seguente: se uno stato mentale m si produce sia quando F_1 è istanziata sia quando F_2 è istanziata (per esempio, perché F_2 è una condizione che causa normalmente la stessa risposta di F_1), allora la relazione di indicazione nomologica vale sia per F_1 sia per $F_1 \vee F_2$. Come si determina quale delle due proprietà — la più specifica F_1 o la più generale $F_1 \vee F_2$ — è il contenuto di m ?

La risposta standard di Dretske è che il contenuto è determinato dalla proprietà indicata nelle condizioni di apprendimento, la fase in cui il sistema acquisisce la funzione segnalante: se m ha acquisito la funzione segnalante di indicare F_1 (e non $F_1 \vee F_2$) durante una fase di apprendimento appropriata, allora F_1 è il contenuto anche se il sistema risponde anche a F_2 per effetto di una generalizzazione erronea. Ma nel contesto della ridondanza, questa risposta crea una difficoltà nuova: la funzione segnalante può essere acquisita per un sistema di strutture ridondanti in una fase di apprendimento, ma la struttura specifica della ridondanza può variare nel corso dello sviluppo e dell'expertise, come si è illustrato con l'esempio del sommelier. La funzione segnalante acquisita nelle condizioni originali di apprendimento non determina univocamente la struttura della ridondanza nelle condizioni di utilizzo esperto, e la struttura della ridondanza nell'utilizzo esperto porta contenuti che non erano presenti nella fase di apprendimento.

Per rendere questo punto formalmente preciso, si consideri la decomposizione PID per due strutture ridondanti X_1 e X_2 rispetto a un obiettivo Y in due fasi dell'esperienza del sommelier. Nelle condizioni iniziali di apprendimento (fase A):

$$I(X_1, X_2 ; Y) = R_A + U_{\{X_1, A\}} + U_{\{X_2, A\}} + S_A$$

dove R_A è la ridondanza, $U_{\{X_1, A\}}$ e $U_{\{X_2, A\}}$ sono le informazioni uniche delle strutture X_1 e X_2 , e S_A è la sinergia. Nelle condizioni di expertise (fase B):

$$I(X_1, X_2 ; Y) = R_B + U_{\{X_1, B\}} + U_{\{X_2, B\}} + S_B$$

con $S_B > S_A$ e $R_B < R_A$. L'expertise modifica la struttura interna della firma informazionale, aumentando la componente sinergica — informazione che emerge solo dall'analisi congiunta delle strutture — a scapito della ridondanza pura. Il contenuto del sistema in fase B include componenti sinergiche che non erano presenti in fase A, e queste componenti non hanno una funzione segnalante acquisita nelle condizioni di apprendimento originale. La teoria dretsiana non ha le risorse per spiegare come queste componenti sinergiche acquisiscono contenuto, a meno di non ammettere che la struttura della ridondanza è essa stessa un portatore di contenuto, il che è la tesi ridondantista.

III.3,5 — Il problema della coerenza interna come condizione di contenuto

La terza difficoltà per la teoria dretsiana è forse la più profonda, e riguarda il ruolo che la coerenza interna tra rappresentazioni ridondanti svolge nella costituzione del contenuto. Come si è mostrato nel Capitolo IV — anticipando qui i risultati formali — la misura di errore rappresentazionale in un sistema ridondante è definita come distanza dalla firma di riferimento:

$$\text{Err}(\Xi, \Theta) = d_GH(\Sigma(\Xi, \Theta), \Sigma^*(\Theta))$$

dove $\Sigma(\Xi, \Theta)$ è la firma informazionale effettiva del sistema Ξ per l'obiettivo Θ , $\Sigma^*(\Theta)$ è la firma di riferimento, e d_GH è la distanza di Gromov-Hausdorff tra le due strutture reticolari. Il fatto cruciale è che questa misura di errore non riguarda la corrispondenza di singole strutture individuali con proprietà del mondo, ma la struttura complessiva delle relazioni tra le rappresentazioni multiple. Un sistema può avere ciascuna delle sue strutture individuali perfettamente calibrata — ciascuna indica nel modo corretto la proprietà che ha la funzione segnalante di indicare — ma il sistema nel suo complesso può avere un contenuto errato se la struttura di coerenza tra le strutture è deviante rispetto alla firma di riferimento.

Questo risultato ha un'importanza filosofica considerevole perché mostra che l'errore rappresentazionale non è riducibile a malfunzionamento di

strutture individuali, e non è nemmeno riducibile a mancanza di corrispondenza tra singole strutture e il mondo. L'errore può consistere in una configurazione sbagliata delle relazioni tra strutture altrimenti corrette. Un caso controesempio illustra questo punto. Immaginiamo un soggetto con prosopagnosia associativa, nel quale il giro fusiforme risponde normalmente ai volti (indica correttamente i pattern fisiognomici), l'amigdala risponde normalmente alla valenza emotiva degli stimoli, ma la connessione tra le due aree è interrotta, cosicché la risposta dell'amigdala a un volto emotivamente familiare non è modulata dall'input del giro fusiforme. In questo caso, la struttura di coerenza tra le rappresentazioni del volto è anormale: la firma informazionale del sistema si discosta dalla firma di riferimento perché manca la componente di ridondanza che collega l'identità visiva del volto alla sua valenza emotiva. Il soggetto non riconosce il volto di sua madre come qualcosa di emotivamente carico, pur riconoscendola visivamente come una specifica identità fisiognomica. La teoria dretsiana direbbe che ciascuna struttura ha il suo contenuto corretto — il giro fusiforme indica correttamente l'identità visiva, l'amigdala risponde correttamente all'input che riceve — ma il contenuto complessivo del sistema è errato perché manca la struttura di coerenza che costituisce il contenuto integrato «volto di mia madre». Questo è un tipo di errore rappresentazionale che la teoria dretsiana non può classificare, perché nessuna delle strutture individuali è mal-calibrata in senso dretsiano: il malfunzionamento è nella struttura delle relazioni, non nei portatori individuali.

III.4 — Il fodorismo e il linguaggio del pensiero: modularità, composizionalità e il nodo centrale

III.4,1 — La teoria fodoriana nella sua doppia articolazione

Il sistema di Jerry Fodor costituisce, per molti aspetti, il punto di riferimento ineludibile per qualsiasi teoria del contenuto mentale sviluppata nella seconda metà del Novecento. Non perché tutte le teorie successive siano fodoriane, ma perché la posizione fodoriana occupa in modo così netto lo spazio delle possibilità che anche le teorie non-fodoriane si definiscono per lo più in relazione a essa. Fodor ha articolato due tesi strettamente connesse ma distinguibili: la tesi della modularità, presentata nella sua forma sistematica in *The Modularity of Mind* del 1983, e la tesi del linguaggio del pensiero, sviluppata in *The Language of Thought* del 1975 e ulteriormente elaborata in *Concepts* del 1998 e *LOT 2* del 2008. Le due tesi si supportano a vicenda ma non si implicano logicamente: è concepibile sostenere la tesi del linguaggio del pensiero senza sostenere la tesi della modularità, o viceversa. Tuttavia,

nella loro combinazione, producono una concezione dell'architettura cognitiva che è particolarmente vulnerabile alle sfide poste dalla ridondanza rappresentazionale.

La tesi della modularità afferma che i sistemi cognitivi periferici — quelli preposti alla percezione sensoriale, al riconoscimento linguistico, alla comprensione della struttura della scena visiva — sono moduli nel senso tecnico di Fodor: sono specifici per dominio, obbligatori nell'attivazione, rapidi, non accessibili introspezione in modo diretto, hanno architettura neurale fissa, mostrano pattern caratteristici di dissociazione patologica, e sono informativamente incapsulati. L'incapsulamento informativo è la proprietà più discussa e teoricamente più pesante: un modulo elabora il suo input sensoriale utilizzando esclusivamente le informazioni interne al modulo, senza accesso alle credenze e alle aspettative del sistema centrale. Questo significa che l'output del modulo è determinato dall'input sensoriale e dalle strutture interne del modulo, non dalle aspettative contestuali del soggetto.

La tesi del linguaggio del pensiero afferma che i processi cognitivi di ordine superiore — ragionamento, formazione delle credenze, pianificazione dell'azione — operano su rappresentazioni mentali che hanno struttura sintattica nel senso della logica formale: sono simboli complessi costruiti da simboli atomici attraverso operazioni di combinazione sintatticamente definite, e le loro proprietà semantiche supervengono sulla loro struttura sintattica secondo regole composizionali. Il ragionamento è elaborazione sintattica che preserva la veridicità perché i simboli del Mentalese sono token di tipi sintatticamente definiti le cui relazioni sintattiche rispecchiano le relazioni semantiche rilevanti per la validità del ragionamento.

III.4,2 — La ridondanza intramodulare e i suoi limiti

All'interno dei moduli sensoriali, la ridondanza gerarchica non è problematica per il fodorismo: è semplicemente la struttura interna dell'elaborazione modulare. Un modulo visivo che elabora l'input attraverso stadi progressivi di analisi — caratteristiche elementari, configurazioni di caratteristiche, pattern gestalici, categorie di oggetti — produce una gerarchia di rappresentazioni ridondanti nel senso che ogni stadio rappresenta il medesimo input a un grado diverso di astrazione. Questa ridondanza intramodulare è compatibile con il principio di incapsulamento: ogni stadio dell'elaborazione è incapsulato rispetto alle informazioni proveniente da altri moduli, anche se i diversi stadi si influenzano gerarchicamente all'interno del modulo.

Il primo problema emerge, tuttavia, quando si considera che il principio di incapsulamento non è empiricamente ben supportato per nessun sistema sensoriale studiato in dettaglio. La corteccia visiva primaria non è incapsulata nel senso di Fodor: riceve feedback discendente massivo dalle aree di ordine superiore, e questo feedback modifica la risposta delle

cellule della corteccia visiva primaria in modo dipendente dalle aspettative e dall'attenzione. Gli esperimenti di Nancy Kanwisher e colleghi con la fMRI adattata hanno mostrato che la risposta delle prime aree visive a stimoli ripetuti dipende dall'aspettativa del soggetto riguardo alla ripetizione: se il soggetto si aspetta un cambiamento e invece lo stimolo è ripetuto, la risposta adattiva è soppressa, il che indica che l'aspettativa modula l'elaborazione a livello delle prime aree visive. Questo risultato è incompatibile con il principio di incapsulamento in senso stretto: l'elaborazione del primo livello visivo non è determinata esclusivamente dall'input sensoriale locale ma è modulata top-down dalle aspettative.

Il fodorismo può rispondere a questi risultati sostenendo che il feedback discendente è un'indicazione di come la modulazione delle aspettative influisce sui meccanismi di attenzione, non sui meccanismi computazionali del modulo sensoriale in senso stretto. L'attenzione è, in questa difesa, una manopola di guadagno che amplifica certi segnali sensoriali rispetto ad altri senza modificare i processi computazionali che trasformano gli input sensoriali in output modulari. Questa difesa è concepibile, ma sposta il problema: i meccanismi di attenzione, se sono sensibili alle credenze e alle aspettative del sistema centrale, non sono essi stessi incapsulati, e la loro influenza sull'elaborazione modulare introduce una forma di penetrabilità cognitiva che il principio di incapsulamento dovrebbe precludere.

III.4,3 — La ridondanza intermodale e il problema del sistema centrale

La difficoltà più grave per il fodorismo emerge nella zona di confine tra i moduli e il sistema centrale, che è precisamente la zona in cui la ridondanza intermodale — la presenza dello stesso contenuto in formati diversi corrispondenti a diversi moduli sensoriali — deve essere integrata in una rappresentazione unitaria nel Mentalese. Fodor è esplicito sul fatto che questa integrazione è il compito del sistema centrale, e altrettanto esplicito sul fatto che il sistema centrale è fondamentalmente opaco alla teoria scientifica: la sua complessità isotropica e quiniana lo rende non studiabile con metodi scientifici di decomposizione modulare.

Questa posizione teorica è, da un certo punto di vista, onesta: Fodor ammette che non sa come il sistema centrale funzioni. Ma dal punto di vista dell'analisi della ridondanza rappresentazionale, questa ammissione è problematica perché localizza il problema esattamente nel luogo in cui la ridondanza costitutiva ha le sue manifestazioni più importanti. Se il sistema centrale è opaco, e se la ridondanza costitutiva opera principalmente nel sistema centrale, allora la teoria fodoriana non ha le risorse per trattare il fenomeno che questa monografia intende analizzare.

Il problema si aggrava considerando la struttura del Mentalese come sistema di simboli con semantica composizionale. Nella semantica composizionale, il contenuto di un simbolo complesso è determinato dai contenuti dei suoi costituenti simbolici e dalla struttura sintattica della loro combinazione. Questo principio è vincolante nel senso che ogni componente dell'informazione trasportata da un simbolo complesso deve essere rintracciabile nei contenuti dei costituenti simbolici o nella struttura combinatoria. Informazione che non è riducibile né ai contenuti dei costituenti né alla struttura combinatoria — informazione sinergica nel senso della PID — non ha posto nella semantica del Mentalese.

Si costruisca un caso controesempio formale. Consideriamo due rappresentazioni nel Mentalese del sistema centrale, R_1 e R_2 , che rispecchiano rispettivamente l'output del modulo visivo e del modulo verbale per il medesimo oggetto O . R_1 porta l'informazione $I(X_1; Y)$ relativa alle caratteristiche visive di O , e R_2 porta l'informazione $I(X_2; Y)$ relativa alle caratteristiche verbali di O . Nella semantica composizionale del Mentalese, il simbolo complesso $R_1 \& R_2$ porta l'informazione $I(X_1; Y) + I(X_2; Y) - I(X_1 \cap X_2; Y)$, ossia la somma delle informazioni meno la ridondanza tra le due sorgenti. Questo è corretto per la componente additiva dell'informazione. Ma la componente sinergica $S(X_1, X_2; Y)$ — l'informazione che emerge solo dall'analisi congiunta di X_1 e X_2 e che non è presente né in X_1 né in X_2 individualmente — non compare in $I(R_1 \& R_2)$. Formalmente:

$$I(X_1, X_2; Y) = I(X_1; Y) + I(X_2; Y) - R(X_1, X_2; Y) + S(X_1, X_2; Y)$$

dove R è la ridondanza e S è la sinergia. Il simbolo composizionale $R_1 \& R_2$ cattura:

$$I(R_1 \& R_2; Y) = I(X_1; Y) + I(X_2; Y) - R(X_1, X_2; Y)$$

ma non cattura $S(X_1, X_2; Y)$. La componente sinergica è, in senso rigoroso, non rappresentata nel Mentalese composizionale. Un sistema che ragiona in Mentalese composizionale perde sistematicamente la componente sinergica del contenuto integrato dei diversi moduli.

Questo risultato formale ha un'interpretazione cognitiva precisa: il Mentalese composizionale è in linea di principio insufficiente per rappresentare i contenuti che emergono dalla fusione sinergica di informazioni multimodali. I contenuti come «questo odore evoca questo colore» (nell'esperienza sinestetica), «questo ritmo corrisponde a questo gesto» (nella percezione audiovisiva del movimento), «questa voce appartiene a questo volto» (nel riconoscimento audiovisivo dell'identità) sono contenuti che non si lasciano catturare dalla composizione dei contenuti dei moduli coinvolti: sono contenuti sinergici che emergono dalla struttura delle relazioni tra i moduli, e che richiederebbero, per essere rappresentati nel Mentalese, simboli primitivi che codificano le relazioni intermodali — simboli che la teoria fodoriana non contempla.

III.4,4 — Il problema dell'olismo semantico e la risposta atomistica

Fodor è esplicitamente atomista nella sua teoria dei concetti, e l'atomismo è la sua risposta al problema dell'olismo semantico. L'olismo semantico — la tesi che il contenuto di ogni concetto dipende dalle sue relazioni inferenziali con tutti gli altri concetti del sistema — è considerato da Fodor inaccettabile per due ragioni: rende impossibile la condivisione di concetti tra individui con sistemi di credenze parzialmente diversi, e rende impossibile il cambiamento di credenza senza cambiamento di contenuto concettuale. La risposta atomistica è che il contenuto di un concetto dipende esclusivamente dalla sua relazione nomologica con la proprietà nel mondo che esso locking: il concetto CANE è agganciato causalmente alle proprietà dei cani, e questo aggancio causale determina il contenuto indipendentemente da qualsiasi relazione inferenziale con altri concetti.

Il problema che la ridondanza costitutiva pone all'atomismo fodoriano è di ordine diverso dall'olismo semantico classico, ma è ugualmente serio. La teoria qui difesa non afferma che il contenuto di ogni rappresentazione dipende dalle sue relazioni inferenziali con tutte le altre: afferma che il contenuto di un sistema di rappresentazioni ridondanti dipende dalla struttura delle relazioni di ridondanza, unicità e sinergia tra le componenti del sistema. Questo non è olismo semantico nel senso classico: non è la tesi che il contenuto dipende dall'intera rete inferenziale. È la tesi, più specifica, che la firma informazionale del sistema — la struttura di dipendenze informatiche tra le componenti ridondanti — è essa stessa parte del contenuto del sistema, e non è riducibile ai contenuti atomistici delle componenti individuali.

Per rendere precisa questa tesi, si consideri la differenza tra due sistemi cognitivi che hanno lo stesso insieme di concetti atomici ma configurazioni diverse della firma informazionale. Il sistema A ha il concetto ROSSO con alta ridondanza tra le sue componenti visiva, verbale e motoria (molti soggetti concordano sulla attribuzione del contenuto ROSSO a stimoli diversi, e le componenti del sistema si corroborano a vicenda). Il sistema B ha lo stesso concetto ROSSO ma con bassa ridondanza tra le componenti — ciascuna componente è relativamente autonoma, le correlazioni tra le componenti visiva, verbale e motoria sono deboli. Nell'atomismo fodoriano, entrambi i sistemi hanno il medesimo concetto ROSSO, con lo stesso contenuto determinato dall'aggancio nomologico con la proprietà dei rossi nel mondo. Ma nella teoria della firma informazionale, i due sistemi hanno concetti di ROSSO con firme diverse: il sistema A ha una firma ad alta ridondanza ($\propto$ alto, ricchezza alta, solidità alta rispetto alle perturbazioni), il sistema B ha una firma a bassa ridondanza ($\propto$ basso, ricchezza bassa, solidità bassa). La teoria atomistica di Fodor non ha le risorse per distinguere questi due casi come casi con contenuti diversi, ma la teoria della firma informazionale afferma che essi corrispondono a strutture

rappresentazionali genuinamente diverse con proprietà semantiche diverse — in particolare, diverse condizioni di errore e diverse generalizzazioni attese a nuovi stimoli.

III.4,5 — Un contro-caso: il conservatorio di flauto e la struttura dell'expertise

Si costruisca un contro-caso particolarmente istruttivo che mostra come la struttura della firma informazionale vari sistematicamente con l'expertise in un modo che la teoria fodoriana non può catturare. Consideriamo due soggetti, un musicista non esperto (M_1) e un flautista professionista (M_2), che percepiscono lo stesso brano eseguito da un flauto traverso. Entrambi hanno il concetto di «flauto» — possono classificare lo strumento come un flauto, ne conoscono l'aspetto visivo, sanno che produce suoni con la bocca applicata a un foro d'imboccatura. Nell'atomismo fodoriano, entrambi i soggetti hanno lo stesso contenuto per il concetto FLAUTO, determinato dall'aggancio nomologico con le proprietà dell'oggetto flauto.

Eppure la struttura della firma informazionale per il concetto FLAUTO è radicalmente diversa nei due soggetti. M_1 ha una firma che include L_1 (il profilo spettrale del suono come caratteristiche fisiche), L_2 (il riconoscimento categoriale dello strumento come flauto), e L_3 (le associazioni autobiografiche con il suono del flauto). M_2 ha una firma molto più ricca: L_1 - L_3 più L_4 (la struttura armonico-timbrica della colonna d'aria nel tubo, modulata dai registri), L_5 (il profilo articolatorio del flautista — il vibrato, lo staccato, la dinamica dello scambio di labbra), L_6 (la posizione nella tradizione esecutiva storica — il flauto barocco versus il flauto moderno Boehm), L_7 (la competenza motoria simulativa nella corteccia premotoria del flautista, che "sente" nell'ascolto la sensazione cinestetica di produrre quei suoni). La firma di M_2 ha una struttura di ridondanza, unicità e sinergia profondamente diversa da quella di M_1 , con componenti sinergiche tra L_4 , L_5 e L_7 che sono assenti nel sistema di M_1 .

La differenza di firma produce differenze genuine nel contenuto: M_2 sente nel brano aspetti che M_1 non sente — la qualità dell'imboccatura del flautista, la precisione delle dita nell'ornamentazione, la calibrazione del respiro nelle frasi lunghe. Questi aspetti non sono semplicemente «più dettagli» dello stesso contenuto atomistico: sono componenti del contenuto che emergono dalla struttura sinergica della firma esperta, e che non sono presenti nella firma di M_1 in nessun livello. Fodor avrebbe probabilmente risposto che M_1 e M_2 hanno concetti diversi di «flauto» — M_2 ha un concetto più ricco con più relazioni inferenziali e una storia di causazione più articolata. Ma questa risposta cede all'olismo che Fodor vuole evitare, perché la differenza tra le firme di M_1 e M_2 non è una differenza atomistica di aggancio causale con il mondo (entrambi sono agganciati causalmente ai flauti) ma una differenza olistica nella struttura delle relazioni tra le componenti del sistema.

III.4,6 — La tesi della composizionalità difensiva e il suo fallimento

Di fronte alle difficoltà accumulate, una versione difensiva del fodorismo potrebbe sostenere che la sinergia informativa non è, in realtà, un esempio di contenuto non composizionale, ma è piuttosto il risultato di un'applicazione incompleta della semantica composizionale. Secondo questa difesa, la sinergia che emerge dall'analisi congiunta di X_1 e X_2 rispetto a Y è il contenuto di un concetto relazionale $R(X_1, X_2)$ che non era stato previsto nella scomposizione atomistica iniziale, ma che è perfettamente composizionale: il suo contenuto è la funzione composizionale del contenuto di R applicato alla coppia (X_1, X_2) . La sinergia non sarebbe in conflitto con la composizionalità ma sarebbe un'espansione del vocabolario del Mentalese per includere i simboli relazionali appropriati.

Questa difesa è ingegnosa ma insoddisfacente per una ragione di principio. La sinergia informazionale in senso PID è irriducibile alla composizione di qualsiasi insieme finito di simboli relazionali aggiunti al vocabolario del Mentalese, perché la struttura delle relazioni sinergiche dipende dall'intera struttura delle dipendenze statistiche tra le componenti del sistema, dipendenze che cambiano continuamente con l'esperienza, con il contesto, e con lo stato di attivazione del sistema. Aggiungere simboli relazionali al vocabolario del Mentalese per catturare la sinergia è un processo che non converge: per ogni insieme finito di simboli relazionali aggiunti, vi saranno strutture sinergiche di ordine superiore che richiedono ulteriori simboli relazionali. Formalmente, la sinergia di ordine n richiede simboli n -ari che non sono riducibili a combinazioni di simboli di ordine inferiore, e il numero di ordini pertinenti non ha un limite superiore fissato a priori. La composizionalità del Mentalese è, in questo senso, strutturalmente inadeguata a catturare la piena struttura del contenuto in un sistema ridondante di rappresentazioni biologiche.

III.5 — Le teorie predittive: la distribuzione come implementazione, non come struttura

Le teorie predittive del tipo di Friston e Clark rappresentano il tentativo più sistematico di integrare la distribuzione delle rappresentazioni nel cervello in una teoria del contenuto coerente. Nell'approccio del Principio dell'Energia Libera, il cervello è un sistema che minimizza l'energia libera variazionale — una misura della discrepanza tra il modello generativo del sistema e l'input sensoriale — attraverso l'aggiornamento continuo delle sue credenze a posteriori. Le rappresentazioni neurali

sono implementazioni fisiche delle distribuzioni di probabilità che compongono il modello generativo; la gerarchia delle aree cerebrali corrisponde alla gerarchia dei livelli del modello generativo, con livelli superiori che predicono i parametri dei livelli inferiori e livelli inferiori che inviano segnali di errore di previsione ai livelli superiori.

Questo framework ha il grande vantaggio di spiegare naturalmente la ridondanza gerarchica: ogni livello della gerarchia è una rappresentazione del medesimo stimolo a un diverso grado di astrazione, e la struttura della gerarchia corrisponde alla struttura del modello generativo. La ridondanza non è un'anomalia ma una caratteristica strutturale del modello gerarchico: ogni livello porta informazione sul medesimo stimolo, e la struttura delle dipendenze tra livelli riflette la struttura del modello generativo del sistema.

Ma la teoria predittiva trova difficoltà specifiche quando si tratta di caratterizzare la sinergia — la componente dell'informazione che emerge solo dall'analisi congiunta di rappresentazioni a livelli diversi. Nel framework dell'energia libera, la distribuzione a posteriori è il risultato dell'integrazione di tutta l'informazione disponibile attraverso la gerarchia; essa cattura, in linea di principio, la struttura dell'informazione congiunta di tutti i livelli. Ma la distribuzione a posteriori è un singolo oggetto matematico — una distribuzione di probabilità sullo spazio degli stati latenti — che è il risultato dell'integrazione, non la struttura dell'integrazione. La struttura delle relazioni tra i diversi livelli gerarchici — le componenti di ridondanza, unicità e sinergia che la PID identifica — è invisibile nella distribuzione a posteriori finale.

Questo è il senso preciso in cui le teorie predittive condividono la presupposizione di unicità del portatore: il portatore del contenuto è la distribuzione a posteriori, non la struttura delle relazioni tra i livelli che producono quella distribuzione. La differenza tra una distribuzione a posteriori prodotta attraverso integrazione di rappresentazioni altamente ridondanti e una prodotta attraverso integrazione di rappresentazioni altamente sinergiche può essere computazionalmente trascurabile — le due distribuzioni a posteriori possono essere identiche — ma la struttura delle relazioni tra le rappresentazioni che le producono è profondamente diversa, e quella differenza è rilevante per il tipo di robustezza, generalizzabilità e vulnerabilità del sistema.

III.6 — Sinergia informativa e fallimento del principio di composizionalità: una trattazione estesa

III.6,1 — La sinergia come fenomeno irreducibile

Il fenomeno della sinergia informativa — la presenza di informazione che emerge solo dall'analisi congiunta di rappresentazioni multiple, senza essere portata da nessuna di esse individualmente — costituisce il punto di massima tensione tra il quadro della ridondanza costitutiva e le teorie del contenuto basate sul principio di composizionalità. È opportuno affrontare la struttura formale di questo fenomeno in modo più sistematico di quanto non sia stato fatto finora, perché è la comprensione precisa della sinergia che permette di valutare correttamente il portato delle obiezioni alle teorie composizionali.

Nella decomposizione PID introdotta da Williams e Beer e sviluppata da Bertschinger e colleghi, per un sistema con n sorgenti $S = \{X_1, \dots, X_n\}$ e obiettivo Y , l'informazione mutua totale $I(X_1, \dots, X_n; Y)$ si decompone in atomi informativi che corrispondono a contribuzioni irriducibili di diversi sottoinsiemi delle sorgenti. L'insieme degli atomi per $n=3$ è indicizzato dal reticolo degli antichain del reticolo degli antichain della struttura di potenza, che per $n=3$ ha 18 elementi. Tra questi elementi vi sono le componenti sinergiche di ordine 2 e di ordine 3:

Per $n=2$, la sinergia è l'unico atomo che corrisponde all'antichain $\{\{X_1, X_2\}\}$, ed è l'informazione su Y che è disponibile congiuntamente in X_1 e X_2 ma non disponibile in nessuno dei due individualmente:

$$S(X_1, X_2; Y) = I(X_1, X_2; Y) - I(X_1; Y) - I(X_2; Y) + R(X_1, X_2; Y)$$

Per $n=3$, oltre alle sinergie di ordine 2 tra ogni coppia, vi è una sinergia di ordine 3 $S_3(X_1, X_2, X_3; Y)$ che è l'informazione disponibile solo nell'analisi congiunta di tutte e tre le sorgenti e non in nessuna combinazione di due.

La sinergia di ordine 3 è l'esempio più netto di informazione strutturalmente non composizionale: non è la funzione di nessuna combinazione di informazioni di ordine inferiore, e non è riducibile a proprietà di singole sorgenti o di coppie. Nella cognizione biologica, le strutture sinergiche di ordine superiore corrispondono a contenuti che emergono solo dalla coordinazione simultanea di molti sistemi rappresentazionali.

III.6,2 — Un caso paradigmatico: la percezione della foresta

Un esempio concreto e costruito con cura di sinergia di ordine superiore nella cognizione biologica è quello che si potrebbe chiamare la percezione della foresta. Un individuo che cammina in una foresta densa percepisce un contenuto ricco e strutturato — la profondità della foresta, l'intrico delle chiome, il gioco di luce tra le foglie, il suono del vento negli alberi, il profumo dell'humus, la sensazione tattile del suolo soffice sotto i piedi — che non è la semplice somma dei percetti di ciascuna modalità sensoriale coinvolta. Il contenuto integrato «foresta» include componenti

sinergiche che emergono solo quando la struttura di coerenza tra le diverse modalità è integra: il suono del vento è percepito come il suono del vento tra quegli alberi — quegli specifici alberi con quelle specifiche caratteristiche visive — e questa percezione integrata non è riducibile alla somma del percepito visivo degli alberi e del percepito uditivo del vento.

Formalmente, sia X_{vis} la sorgente visiva (caratteristiche spaziali della foresta), X_{aud} la sorgente uditiva (caratteristiche acustico-spaziali del vento), X_{olv} la sorgente olfattiva (profilo chimico dell'ambiente forestale), X_{tatt} la sorgente tattile (caratteristiche meccaniche del terreno), e Y il contenuto integrato «foresta». La firma informativa di questo percepito include, per ogni coppia di sorgenti, componenti di ridondanza e sinergia. Ma include anche la sinergia di ordine 4:

$$S_4(X_{vis}, X_{aud}, X_{olv}, X_{tatt} ; Y) > 0$$

Questa componente è l'informazione sul contenuto «foresta» che emerge solo dall'analisi congiunta di tutte e quattro le sorgenti simultaneamente, e non da nessuna combinazione di tre o meno. Essa corrisponde, fenomenologicamente, a ciò che si potrebbe descrivere come la qualità dell'immersione nella foresta — il senso di trovarsi in un ambiente coerente e completo — che non si riduce a nessuna delle sue componenti sensoriali e nemmeno alla loro somma.

Questa struttura sinergica di ordine superiore non ha posto in nessuna delle teorie esaminate. La teleosemantica non ha le risorse per attribuire una funzione propria alla sinergia di ordine superiore (quale dispositivo la produce, quale dispositivo la consuma?). La teoria dretsiana non ha i mezzi per identificare la proprietà del mondo che S_4 indica (quale proprietà nomologica corrisponde all'informazione che è presente solo nell'analisi congiunta di quattro sorgenti?). Il fodorismo non può rappresentare S_4 in Mentalese compositivo. Le teorie predittive integrano la sinergia nella distribuzione a posteriori senza tematizzarla come struttura del contenuto.

III.6,3 — Implicazioni per la semantica delle lingue naturali

Il fallimento del principio di composizionalità per i contenuti mentali sinergici ha implicazioni che si estendono oltre la filosofia della mente verso la semantica delle lingue naturali. La semantica compositiva delle lingue naturali è costruita sulla premessa che il significato delle espressioni complesse sia una funzione del significato delle loro parti e della struttura della loro combinazione. Se i contenuti mentali che le espressioni linguistiche esprimono includono componenti sinergiche irriducibili, allora la semantica compositiva può catturare al più la componente compositiva di quel contenuto, non la componente sinergica.

Questo risultato ha una forma più precisa che vale la pena articolare. Sia $C(E)$ il contenuto mentale associato all'espressione linguistica E , e sia $C_{\text{comp}}(E)$ la componente di $C(E)$ che è catturabile da una semantica composizionale. Allora $C_{\text{comp}}(E) \neq C(E)$ in tutti i casi in cui $C(E)$ include componenti sinergiche. La differenza $C(E) - C_{\text{comp}}(E)$ è la componente del contenuto mentale che rimane inesprimibile attraverso la semantica composizionale dell'espressione: è il surplus semantico delle espressioni linguistiche, il contenuto che l'espressione evoca ma non può letteralmente portare in virtù della sua struttura composizionale.

Questa idea non è nuova — è presente in forme diverse nella fenomenologia (la differenza tra significato intenzione e significato riempimento in Husserl), nella poetica (la differenza tra denotazione e connotazione nelle figure poetiche), e nella pragmatica (la differenza tra contenuto semantico e contenuto comunicativo in Grice). Ma il quadro della ridondanza costitutiva fornisce, per la prima volta, una caratterizzazione formale precisa di questo surplus semantico come componente sinergica della firma informazionale, e lo colloca all'interno di una teoria del contenuto mentale che spiega come questa componente emerge e quali sono le sue condizioni di variazione.

III.7 — Verso un quadro alternativo: la firma informazionale come risposta alle insufficienze

L'analisi condotta nelle sezioni precedenti converge su una diagnosi comune: ogni teoria esaminata incontra difficoltà sistematiche nella zona della ridondanza costitutiva perché condivide la presupposizione di unicità del portatore del contenuto. La teleosemantica individua il portatore nella struttura che ha la funzione propria rilevante; la teoria dretsiskiana individua il portatore nella struttura che ha la funzione segnalante; il fodorismo individua il portatore nel simbolo del Mentalese; le teorie predittive individuano il portatore nella distribuzione a posteriori. In tutti i casi, il contenuto è associato a un singolo oggetto matematico, anche se questo oggetto può essere complesso o distribuito: è l'oggetto che determina il contenuto, e la struttura delle relazioni tra le rappresentazioni multiple è, al più, rilevante per la determinazione dell'oggetto ma non è essa stessa portatrice di contenuto.

La proposta alternativa — elaborata in dettaglio nel Capitolo IV — è che il portatore del contenuto non sia un singolo oggetto matematico ma una struttura relazionale: la firma informazionale del sistema, definita come il reticolo degli atomi PID che decompone l'informazione congiunta del sistema rispetto all'obiettivo. La firma informazionale non è un singolo vettore o distribuzione di probabilità: è una struttura di ordine parziale

— un reticolo — che codifica le dipendenze informatiche tra le componenti del sistema a tutti gli ordini. Due sistemi con la stessa firma informazionale hanno lo stesso contenuto, indipendentemente dalla struttura fisica dei loro portatori individuali; due sistemi con firme diverse hanno contenuti diversi, anche se ogni singolo portatore individuale ha la stessa struttura e le stesse proprietà.

Questa proposta risponde alle obiezioni accumulate in modo strutturale. Il problema del livello di descrizione per Millikan si dissolve perché la firma informazionale non è definita a un livello fisico specifico: è definita dalla struttura delle dipendenze statistiche tra le componenti, e questa struttura è invariante rispetto a molte trasformazioni fisiche. Il problema del disjunction per Dretske si dissolve perché la firma non è determinata dall'indicazione di singole strutture ma dalla struttura globale delle dipendenze; sistemi che producono lo stesso output in condizioni diverse possono avere firme diverse se la struttura delle loro dipendenze interne è diversa. Il problema della composizionalità per Fodor si dissolve perché la firma include esplicitamente le componenti sinergiche di ordine superiore, che non sono catturabili dalla composizione dei contenuti delle componenti individuali. Il problema del portatore singolo per le teorie predittive si dissolve perché la firma è la struttura dell'integrazione, non il risultato dell'integrazione.

Capitolo IV — Verso una teoria formale: la decomposizione dell'informazione parziale e la firma del contenuto

IV.1 — Il problema della misurazione del contenuto distribuito

Il compito di questo capitolo è costruire il quadro formale che permette di caratterizzare il contenuto di un sistema di rappresentazioni ridondanti in modo preciso e misurabile. Il punto di partenza è il riconoscimento che qualsiasi teoria del contenuto distribuito deve rispondere a due esigenze che sono in tensione tra loro: da una parte, deve permettere di identificare il contenuto come qualcosa di determinato — altrimenti la nozione di contenuto perde il suo mordente esplicativo; dall'altra, deve permettere di caratterizzare la struttura della molteplicità delle rappresentazioni — non solo il fatto che vi siano molte rappresentazioni, ma come esse si relazionano tra loro e come questa struttura di relazione costituisce il contenuto.

Lo strumento formale che permette di soddisfare entrambe queste esigenze è la teoria della decomposizione dell'informazione parziale (Partial Information Decomposition, PID), sviluppata da Williams e Beer nel 2010 e successivamente affinata da Griffith, Bertschinger, Rauh e collaboratori. La PID non è stata originariamente sviluppata per la filosofia del contenuto mentale: è uno strumento della teoria dell'informazione che risponde alla domanda «come si decompone l'informazione che un insieme di sorgenti porta su un obiettivo nelle sue componenti elementari irriducibili?». Che questo strumento si riveli illuminante per la filosofia del contenuto non è una coincidenza: riflette il fatto che il problema della struttura dell'informazione nelle reti di sorgenti multiple è, a un certo livello di astrazione, la stessa struttura del problema della molteplicità rappresentazionale.

IV.2 — La struttura reticolare degli atomi informativi

Siano $X_1, X_2, \dots, X$ un insieme di sorgenti di informazione — che nella nostra applicazione corrispondono alle rappresentazioni ridondanti di un sistema cognitivo — e sia Y l'obiettivo — che nella nostra applicazione corrisponde al contenuto che il sistema rappresenta. L'informazione mutua $I(\{X_1, \dots, X\}; Y)$ misura la quantità totale di informazione che l'insieme delle sorgenti porta sull'obiettivo. Il punto di partenza della PID è l'osservazione che questa quantità aggregata non cattura la struttura delle relazioni tra le sorgenti: due sistemi di sorgenti possono avere la stessa informazione mutua totale ma strutture di relazione completamente diverse.

La decomposizione dell'informazione parziale scompone $I(\{X_1, \dots, X\}; Y)$ in atomi irriducibili, ciascuno dei quali corrisponde a una specifica struttura di ridondanza-sinergia tra sottoinsiemi di sorgenti. Il reticolo di questi atomi è costruito sulle anti-chain del reticolo delle partizioni di $\{1, 2, \dots, n\}$: un'anti-chain α è un insieme di sottoinsiemi di $\{1, \dots, n\}$ tale che nessun elemento di α è un sottoinsieme di un altro. Intuitivamente, un'anti-chain $\alpha = \{A_1, A_2, \dots, A\}$ rappresenta la struttura «l'informazione è fornita ridondantemente da tutti i sottoinsiemi $A_1, A_2, \dots, A$, ma non da nessun sottoinsieme proprio di questi». Denotiamo con $\mathcal{A}$ l'insieme di tutte le anti-chain del reticolo delle partizioni di $\{1, \dots, n\}$, e con $\mathcal{R}(\alpha; Y)$ il valore dell'atomo informativo corrispondente all'anti-chain α rispetto all'obiettivo Y .

Per $n = 2$, l'insieme $\mathcal{A}_2$ contiene esattamente quattro anti-chain non banali: $\{1\} \wedge \{2\}$ — la ridondanza bilaterale, $\{1\}$ — l'informazione unica della sorgente 1, $\{2\}$ — l'informazione unica della sorgente 2, $\{1, 2\}$ — la sinergia bilaterale.

La decomposizione dell'informazione mutua $I(\{X_1, X_2\}; Y)$ nei quattro atomi corrispondenti è:

$$I(\{X_1, X_2\}; Y) = \mathcal{R}(\{1\} \wedge \{2\}; Y) + U_1(Y) + U_2(Y) + \text{Syn}(\{1,2\}; Y)$$

dove la ridondanza $\mathcal{R}(\{1\} \wedge \{2\}; Y)$ è l'informazione che sia X_1 sia X_2 forniscono individualmente su Y ; l'informazione unica $U_1(Y)$ è l'informazione che X_1 fornisce su Y ma X_2 non fornisce; $U_2(Y)$ è simmetricamente l'informazione unica di X_2 ; e la sinergia $\text{Syn}(\{1,2\}; Y)$ è l'informazione su Y che emerge solo dall'analisi congiunta di X_1 e X_2 .

La relazione additiva tra questi atomi è:

$$I(X_i; Y) = \mathcal{R}(\{1\} \wedge \{2\}; Y) + U_i(Y) \text{ per } i = 1, 2$$

$$I(\{X_1, X_2\}; Y) = \mathcal{R}(\{1\} \wedge \{2\}; Y) + U_1(Y) + U_2(Y) + \text{Syn}(\{1,2\}; Y)$$

Da queste relazioni segue che la sinergia è definita implicitamente da:

$$\text{Syn}(\{1,2\}; Y) = I(\{X_1, X_2\}; Y) - I(X_1; Y) - I(X_2; Y) + \mathcal{R}(\{1\} \wedge \{2\}; Y)$$

Il termine $I(\{X_1, X_2\}; Y) - I(X_1; Y) - I(X_2; Y)$ è noto come co-informazione $C(X_1; X_2; Y)$ e può essere negativo (segnalando sinergia netta tra le sorgenti) o positivo (segnalando ridondanza netta). La relazione tra co-informazione e atomi PID mostra che la co-informazione è una misura aggregata che non distingue tra ridondanza e sinergia: due sistemi con la stessa co-informazione possono avere strutture di ridondanza e sinergia molto diverse.

Per $n = 3$, la struttura si fa più ricca. Le anti-chain di $\{1, 2, 3\}$ sono 18 elementi non banali, e la decomposizione PID produce 18 atomi irriducibili. Tra questi vi sono atomi che corrispondono a ridondanze di ordine diverso: $\mathcal{R}(\{1\} \wedge \{2\} \wedge \{3\}; Y)$ — la ridondanza trilaterale, l'informazione che tutte e tre le sorgenti forniscono individualmente su Y — $\mathcal{R}(\{1\} \wedge \{2,3\}; Y)$ — la ridondanza tra la sorgente 1 e l'insieme $\{2,3\}$ — e così via per tutte le anti-chain. Le sinergie di ordine 2 e di ordine 3 compaiono anch'esse come atomi distinti nel reticolo.

La struttura del reticolo delle anti-chain è quella di un reticolo parzialmente ordinato: esiste una relazione d'ordine $\leq$ su $\mathcal{A}$ tale che $\alpha \leq \beta$ se ogni elemento di α è contenuto in qualche elemento di β . Questa struttura d'ordine permette di definire la somma cumulativa degli atomi come:

$$I_{\alpha}(Y) \equiv \sum_{\{\beta \leq \alpha\}} \mathcal{R}(\beta; Y)$$

che è l'informazione mutua dell'anti-chain α rispetto all'obiettivo Y nel senso di Williams e Beer. La funzione $\alpha \mapsto I_{\alpha}(Y)$ è monotona crescente rispetto all'ordine parziale su $\mathcal{A}$, il che riflette il fatto che aggiungere sorgenti non può ridurre l'informazione sull'obiettivo.

IV.3 — La firma informazionale: definizione e proprietà

Sulla base della struttura reticolare degli atomi PID, si può ora definire il concetto centrale del presente quadro formale: la firma informazionale di un sistema di rappresentazioni ridondanti.

Sia $\Xi = (\xi_1, \xi_2, \dots, \xi_n)$ un vettore di rappresentazioni ridondanti — le n componenti del sistema cognitivo che concernono il medesimo contenuto — e sia Θ la variabile che rappresenta il contenuto dell'oggetto o la proprietà del mondo che il sistema rappresenta. La firma informazionale di Ξ rispetto a Θ è la tripla:

$$\Sigma(\Xi, \Theta) = (\mathcal{R}(\Xi, \Theta), \kappa(\Xi, \Theta), \sigma(\Xi, \Theta))$$

dove i tre componenti sono definiti come segue.

Il reticolo degli atomi $\mathcal{R}(\Xi, \Theta)$ è l'insieme degli atomi della decomposizione PID di $I(\xi_1, \dots, \xi_n; \Theta)$, cioè l'insieme $\{\mathcal{R}(\alpha; \Theta) : \alpha \in \mathcal{A}\}$. Questo oggetto è un vettore di dimensione $|\mathcal{A}|$ i cui componenti sono i valori degli atomi informativi. Esso cattura la struttura completa delle relazioni di ridondanza, unicità e sinergia tra le rappresentazioni ridondanti rispetto al contenuto.

Il coefficiente di coerenza $\kappa(\Xi, \Theta)$ è definito come:

$$\kappa(\Xi, \Theta) = I_{\cup}(\{1\}, \{2\}, \dots, \{n\}; \Theta) / ((1/n) \cdot \sum_i I(\xi_i; \Theta))$$

dove $I_{\cup}(\{1\}, \{2\}, \dots, \{n\}; \Theta) \equiv \sum_i I(\xi_i; \Theta) - I(\xi_1, \dots, \xi_n; \Theta) + I_{\text{joint}}$ è la somma delle informazioni individuali meno le ridondanze bilaterali — o, equivalentemente, la somma degli atomi di informazione unica e delle informazioni ridondanti di primo ordine. Il coefficiente κ misura il rapporto tra la ridondanza di primo ordine aggregata e l'informazione individuale media: è 1 quando tutte le sorgenti portano esattamente la stessa informazione su Θ (ridondanza pura), e tende a 0 quando le sorgenti portano informazioni completamente disgiunte su Θ .

Il rapporto sinergico $\sigma(\Xi, \Theta)$ è definito come:

$$\sigma(\Xi, \Theta) = \text{Syn}(\{1, \dots, n\}; \Theta) / I(\xi_1, \dots, \xi_n; \Theta)$$

dove $\text{Syn}(\{1, \dots, n\}; \Theta)$ è la sinergia di ordine n — la componente dell'informazione mutua congiunta che non è presente in nessun sottoinsieme proprio delle sorgenti. Il rapporto σ misura la frazione dell'informazione congiunta che è di natura sinergica: è 0 quando tutta l'informazione congiunta è decomponibile come somma delle informazioni delle singole sorgenti, e tende a 1 quando quasi tutta l'informazione emerge soltanto dalla considerazione congiunta dell'intero insieme delle sorgenti.

La firma informazionale $(\mathcal{R}(\Xi, \Theta), \kappa(\Xi, \Theta), \sigma(\Xi, \Theta))$ è un oggetto matematico che vive in uno spazio di dimensione $|\mathcal{A}| + 2$. Due sistemi di rappresentazioni ridondanti che hanno la stessa firma informazionale rispetto a un contenuto Θ sono, in un senso preciso, sistemi che rappresentano Θ «allo stesso modo»: la struttura delle relazioni informative tra le loro componenti rispetto a Θ è identica. Due sistemi con firme diverse rappresentano Θ «in modo diverso», e la distanza tra le firme — che si definirà nella prossima sezione — misura quanto diversamente i due sistemi rappresentano lo stesso contenuto.

Prima di procedere alla definizione di distanza, è utile illustrare il significato qualitativo di diverse strutture della firma informazionale. Un sistema con κ alto e σ basso è un sistema in cui molte rappresentazioni ridondanti portano la medesima informazione sul contenuto, senza che emerga informazione nuova dalla loro considerazione congiunta: si tratta di un sistema di rappresentazioni ridondanti nel senso classico, dove la ridondanza serve principalmente a garantire la robustezza contro il rumore. Un sistema con κ basso e σ alto è, al contrario, un sistema in cui le rappresentazioni portano informazioni molto diverse sul contenuto, e la struttura della loro coerenza congiunta porta informazione aggiuntiva che non è presente in nessuna rappresentazione individuale: si tratta di un sistema dove la molteplicità delle rappresentazioni serve non alla robustezza ma all'arricchimento del contenuto attraverso la sinergia. Il sistema cognitivo umano, come si argomenterà nella sezione seguente, ha una firma informazionale che combina valori significativi sia di κ che di σ : è un sistema sia robusto — le informazioni ridondanti garantiscono la stabilità del contenuto — sia sinergico — le informazioni eterogenee si combinano per costituire un contenuto più ricco di qualsiasi componente individuale.

IV.4 — La definizione operativa degli atomi PID: il problema della misura unica e le sue soluzioni

IV.4,1 — Il problema dell'unicità della misura

La decomposizione PID non è un costrutto unico nel senso matematico che la teoria della misura intende: è un programma di ricerca che ha prodotto molteplici proposte per la misura della ridondanza $\mathcal{R}(\alpha; \Theta)$, e le diverse proposte, pur soddisfacendo lo stesso sistema di assiomi di base, producono valori numericamente diversi in casi non banali. Questa situazione è diversa da quella della teoria di Shannon, dove l'entropia $H(X)$ è univocamente determinata dall'assioma di continuità combinato con l'assioma di additività delle componenti equiprobabili. Per la PID, l'assioma fondamentale di non-negatività non è sufficiente a determinare univocamente la misura: esistono molteplici misure non-negative compatibili con tutti gli assiomi proposti.

Il problema dell'unicità della misura ha due dimensioni. La prima dimensione è analitica: si tratta di determinare quali assiomi aggiuntivi, al di là della non-negatività, restringerebbero lo spazio delle misure ammissibili fino a un punto fisso. La seconda dimensione è applicativa: si tratta di determinare quale misura produce le previsioni più accurate sulla struttura empirica delle dipendenze informatiche nei sistemi neurali, indipendentemente dalle considerazioni assiomatiche. Queste due dimensioni non si sovrappongono necessariamente: la misura assiomaticamente più semplice non è detto che sia quella empiricamente più accurata.

IV.4,2 — La misura $I_{\min}$ di Williams-Beer: derivazione e proprietà

La misura originale proposta da Williams e Beer nel 2010 — che definiamo $I_{\min}$ — è costruita a partire dall'intuizione che la ridondanza tra due sorgenti X_1 e X_2 rispetto all'obiettivo Y è la quantità di informazione su Y che è disponibile da ciascuna sorgente individualmente. Due sorgenti sono ridondanti se ciascuna «sa» qualcosa che anche l'altra sa, e la ridondanza è quanto fanno in comune. Il modo più conservativo di misurare questa quantità è prendere il minimo delle informazioni bilaterali:

$$I_{\min}(\{1\}; Y) = I(X_1; Y) \quad I_{\min}(\{2\}; Y) = I(X_2; Y) \quad I_{\min}(\{1\} \wedge \{2\}; Y) = \min(I(X_1; Y), I(X_2; Y))$$

La generalizzazione a sottoinsiemi di più sorgenti si ottiene ricorsivamente: la ridondanza dell'antichain $\alpha = \{A_1, \dots, A\}$ è il minimo della ridondanza di ciascun elemento A_i , dove la ridondanza di A_i (un sottoinsieme di sorgenti) è $I_{\min}$ dell'antichain formata da A_i da solo:

$$I_{\min}(\alpha; Y) = \min_{A_i \in \alpha} I_{\min}(\{A_i\}; Y)$$

Per la coppia binaria, questo produce la decomposizione:

$$I(X_1, X_2; Y) = \min(I(X_1; Y), I(X_2; Y)) + [I(X_1; Y) - \min(I(X_1; Y), I(X_2; Y))] + [I(X_2; Y) - \min(I(X_1; Y), I(X_2; Y))] + [I(X_1, X_2; Y) - I(X_1; Y) - I(X_2; Y) + \min(I(X_1; Y), I(X_2; Y))] = R + U_1 + U_2 + S$$

dove: $R = \min(I(X_1; Y), I(X_2; Y))$ $U_1 = I(X_1; Y) - R = \max(I(X_1; Y), I(X_2; Y)) - I(X_2; Y)$ se $I(X_1; Y) > I(X_2; Y)$, altrimenti 0 $U_2 = I(X_2; Y) - R = \max(I(X_1; Y), I(X_2; Y)) - I(X_1; Y)$ se $I(X_2; Y) > I(X_1; Y)$, altrimenti 0 $S = I(X_1, X_2; Y) - \max(I(X_1; Y), I(X_2; Y))$

Le proprietà della misura $I_{\min}$ sono:

Monotonicità rispetto all'ordine del reticolo: se $\alpha \leq \beta$ nel reticolo delle antichain, allora $I_{\min}(\beta; Y) \geq I_{\min}(\alpha; Y)$. Questa proprietà garantisce che aggiungere sorgenti non riduce l'informazione.

Simmetria: $I_{\min}$ è invariante sotto permutazioni delle sorgenti.

Normalizzazione: $I_{\min}(\{i\}; Y) = I(X_i; Y)$ per ogni singola sorgente i .

Tuttavia, $I_{\min}$ viola la non-negatività per sistemi con $n \geq 3$ sorgenti in certi casi degeneri. Si consideri il sistema XOR generalizzato: X_1 e X_2 variabili uniformi su $\{0,1\}$, $X_3 = X_1 \text{ XOR } X_2$, $Y = X_1$. In questo sistema, $I(X_1; Y) = 1$ bit, $I(X_2; Y) = 0$ bit, $I(X_3; Y) = 0$ bit. La ridondanza trilaterale $I_{\min}(\{1\} \wedge \{2\} \wedge \{3\}; Y) = \min(1, 0, 0) = 0$. Ma $I_{\min}(\{1\} \wedge \{2,3\}; Y) = \min(I_{\min}(\{1\}; Y), I_{\min}(\{2,3\}; Y)) = \min(1, I(X_2, X_3; Y)) = \min(1, I(X_2, X_3; X_1))$. Per il sistema XOR, $I(X_2, X_3; X_1) = 1$ bit (X_1 è completamente determinato da X_2 e X_3). Quindi $I_{\min}(\{1\} \wedge \{2,3\}; Y) = \min(1, 1) = 1$. Ma $I_{\min}(\{1\}; Y) = 1$. La ridondanza tra $\{1\}$ e $\{2,3\}$ è pari all'informazione di X_1 da solo, il che implica che l'informazione unica di X_1 è zero. Questo produce il risultato che la sinergia nella tripla è negativa: non è possibile assegnare valori non-negativi a tutti gli atomi del reticolo tri-sorgente soddisfacendo la consistenza marginale. L'atomo negatività di $I_{\min}$ per $n=3$ è un problema strutturale che non è eliminabile con aggiustamenti locali della formula.

IV.4,3 — La misura I_{BROJA} : derivazione tramite ottimizzazione convessa

La misura I_{BROJA} di Bertschinger e colleghi (2014) risolve il problema della non-negatività attraverso un approccio completamente diverso. Invece di definire la ridondanza direttamente in termini di informazioni mutue marginali, I_{BROJA} è definita come la soluzione di un problema di ottimizzazione convessa.

L'idea fondamentale è che la ridondanza tra due sorgenti X_1 e X_2 rispetto a Y sia la quantità di informazione su Y che è determinata dalla struttura marginale del sistema — la struttura delle distribuzioni bilaterali $P(X_1, Y)$ e $P(X_2, Y)$ — indipendentemente dalla struttura di dipendenza statistica tra X_1 e X_2 . Due sorgenti X_1 e X_2 forniscono informazione ridondante su Y se quella informazione sarebbe presente anche in un sistema alternativo in cui X_1 e X_2 fossero statisticamente indipendenti ma avessero le stesse distribuzioni marginali con Y .

Formalmente, sia $\Delta(P)$ l'insieme delle distribuzioni di probabilità Q su (X_1, X_2, Y) che hanno le stesse distribuzioni marginali bilaterali di P :

$$\Delta(P) = \{Q : Q(X_1, Y) = P(X_1, Y), Q(X_2, Y) = P(X_2, Y)\}$$

La ridondanza I_{BROJA} è il massimo dell'informazione mutua congiunta tra le due sorgenti e l'obiettivo su questa famiglia di distribuzioni:

$$I_{\text{BROJA}}(\{1\} \wedge \{2\}; Y) = \max_{\{Q \in \Delta(P)\}} I_Q(X_1; Y) + I_Q(X_2; Y) - I_Q(X_1, X_2; Y)$$

dove il pedice Q indica che l'informazione mutua è calcolata rispetto alla distribuzione Q . Questa quantità massimizza la parte dell'informazione congiunta che è attribuibile alle strutture marginali bilaterali, escludendo la struttura di dipendenza statistica di ordine superiore tra le sorgenti.

Il problema di massimizzazione è convesso: l'insieme $\Delta(P)$ è convesso (è l'intersezione di due vincoli affini sulle distribuzioni marginali) e la funzione obiettivo è concava rispetto a Q (per la concavità dell'entropia). Quindi il massimo è unico e numericamente calcolabile attraverso algoritmi di ottimizzazione convessa standard.

Le proprietà di I_BROJA sono:

Non-negatività: $I_BROJA \geq 0$ per qualsiasi distribuzione P e qualsiasi numero di sorgenti n .

Consistenza con i marginali: $I_BROJA(\{i\}; Y) = I(X_i; Y)$ per ogni singola sorgente i .

Monotonicità: se $\alpha \leq \beta$, allora $I_BROJA(\beta; Y) \geq I_BROJA(\alpha; Y)$.

Tuttavia, I_BROJA ha una proprietà controintuitiva che è stata identificata come problematica da Griffith e Koch: la ridondanza di due sorgenti statisticamente indipendenti dato l'obiettivo può essere positiva. Consideriamo due sorgenti $X_1 = Y + \varepsilon_1$ e $X_2 = Y + \varepsilon_2$ con $\varepsilon_1, \varepsilon_2$ indipendenti da Y e tra loro. In questo caso, X_1 e X_2 sono condizionalmente indipendenti dato Y , ma $I_BROJA(\{1\} \wedge \{2\}; Y) = I(X_1; Y) = I(X_2; Y) > 0$. Questo risultato è controintuitivo se si interpreta la ridondanza come «informazione condivisa»: se X_1 e X_2 sono condizionalmente indipendenti dato Y , la loro dipendenza è completamente mediata da Y , e l'intuizione vorrebbe che questa mediazione completa produca ridondanza massima. I_BROJA cattura questa intuizione correttamente. Ma per altri scenari in cui la dipendenza tra le sorgenti è parzialmente mediata da Y e parzialmente da strutture di dipendenza dirette, I_BROJA produce valori che non sempre corrispondono alle intuizioni.

IV.4,4 — La misura I_sx di Griffith-Koch: l'approccio dalla sinergia

Griffith e Koch (2014) hanno proposto di definire la PID dall'altro estremo: invece di definire prima la ridondanza e derivarne la sinergia, si definisce prima la sinergia e si deriva la ridondanza. La motivazione è che la sinergia ha una definizione intuitiva più chiara della ridondanza: la sinergia è l'informazione che emerge dall'analisi congiunta e non è riducibile alle informazioni individuali.

La sinergia di ordine n nel framework di Griffith-Koch è definita come:

$$\text{Syn}_{\{\text{GK}\}}(X_1, \dots, X; Y) = I(X_1, \dots, X; Y) - \max_{A \subset \{1, \dots, n\}} I(X_A; Y)$$

dove X_A denota l'insieme delle sorgenti indicizzate da A . Questa è la differenza tra l'informazione mutua totale e il massimo dell'informazione mutua di qualsiasi sottoinsieme proprio: è la componente dell'informazione congiunta che non è catturata da nessun sottoinsieme delle sorgenti.

Questa definizione è semplice e computazionalmente accessibile (richiede $O(2^n)$ calcoli di informazione mutua, che è ancora esponenziale in n ma non così rapido come il numero di Dedekind). Ha il vantaggio di essere sempre non-negativa (come differenza tra il massimo e il valore ottenuto al massimo). Ma ha lo svantaggio di non distribuire l'informazione in atomi irriducibili nel senso pieno della PID: la sinergia di Griffith-Koch è una misura globale della non-riducibilità, non una struttura reticolare completa di atomi.

IV.4,5 — Confronto tra misure e implicazioni per la firma informazionale

Il confronto tra $I_{\min}$, I_{BROJA} e I_{sx} rivela che non esiste una scelta dominante: ciascuna misura ha vantaggi in certi scenari e svantaggi in altri, e la scelta della misura appropriata dipende dal tipo di struttura informazionale che si intende caratterizzare.

Per la firma informazionale come strumento della teoria del contenuto, le implicazioni sono le seguenti.

Se si usa $I_{\min}$, la firma cattura le componenti di ridondanza nel senso conservativo: solo la ridondanza che è presente in ogni sorgente individuale viene conteggiata come ridondanza. Sistemi in cui alcune sorgenti portano informazione ridondante e altre portano informazione sinergica producono firme con alta sinergia residua. Questo favorisce la diagnosi di strutture sinergiche ma può sottostimare la ridondanza in sistemi con più di 2 sorgenti.

Se si usa I_{BROJA} , la firma cattura le componenti di ridondanza nel senso delle distribuzioni marginali: la ridondanza è la parte dell'informazione congiunta che è attribuibile alle strutture bilaterali, indipendentemente dalle dipendenze di ordine superiore. Questo produce firme con ridondanza più alta che $I_{\min}$ nei casi in cui le dipendenze di ordine superiore amplificano la ridondanza bilaterale.

Per gli scopi della teoria del contenuto, la scelta tra $I_{\min}$ e I_{BROJA} corrisponde a una scelta interpretativa: si intende per ridondanza la coesione informativa conservativa (ogni componente porta individualmente la stessa informazione) o la coesione informativa costruttiva (la struttura delle dipendenze bilaterali produce informazione

condivisa)? Entrambe le interpretazioni sono difendibili in contesti diversi. Per i sistemi neurobiologici in cui le sorgenti sono definite come aree cerebrali che elaborano lo stesso stimolo attraverso pathway diversi, la ridondanza costruttiva di I_{BROJA} sembra più appropriata: la ridondanza emerge dalla struttura delle connessioni tra le aree, non solo dalle loro risposte individuali. Per i sistemi artificiali in cui le sorgenti sono definite come canali di trasmissione progettati per portare informazione identica, la ridondanza conservativa di I_{min} sembra più appropriata.

IV.5 — Coerenza come attrattore: la dinamica delle rappresentazioni ridondanti

IV.5,1 — Dal calcolo statico alla dinamica: la necessità dell'approccio dinamico

Le sezioni precedenti hanno trattato la firma informazionale come un oggetto statico — una struttura di dipendenze informatiche calcolata a partire dalla distribuzione di probabilità congiunta delle sorgenti e dell'obiettivo. Ma un sistema cognitivo biologico non opera in uno stato statico: le sue rappresentazioni si aggiornano continuamente in risposta a nuovi input, le dipendenze informatiche tra le sorgenti cambiano nel corso dell'elaborazione, e la firma risultante varia nel tempo. Per una teoria del contenuto che si proponga di spiegare non solo il contenuto in uno stato stazionario ma anche il processo di formazione del contenuto, l'elaborazione dei dati ambigui, e la resistenza alla perturbazione, è necessaria una teoria dinamica della firma informazionale.

La teoria dinamica che si sviluppa in questa sezione è costruita attorno al concetto di attrattore della dinamica delle rappresentazioni — lo stato verso cui il sistema converge in assenza di perturbazioni esterne, e che corrisponde al contenuto «normale» del sistema in quella configurazione. L'attrattore è un concetto dinamico: non è uno stato fisso nel senso delle reti di Hopfield classiche, ma un insieme invariante dello spazio degli stati verso cui le traiettorie del sistema convergono.

IV.5,2 — La funzione di energia e la sua costruzione

La funzione di energia $V(\Xi)$ è una funzione scalare sullo spazio degli stati del sistema cognitivo Ξ che misura il grado di incoerenza delle rappresentazioni: V è alta quando le rappresentazioni sono mutuamente incoerenti e bassa quando sono mutuamente coerenti. La dinamica del sistema è data dalla discesa del gradiente di V perturbata da rumore stocastico:

$$d\Xi/dt = -\nabla V(\Xi) + \eta(t)$$

dove $\eta(t)$ è un termine di rumore bianco con intensità σ^2 . I minimi locali di V corrispondono agli attrattori della dinamica e, nella tesi della ridondanza costitutiva, al contenuto del sistema.

La costruzione esplicita della funzione V è il punto più delicato della teoria dinamica, perché V deve essere costruita in modo che i suoi minimi corrispondano alle configurazioni di massima coerenza delle rappresentazioni, e la coerenza è definita in termini della firma informazionale — un oggetto che dipende dalle statistiche di secondo e ordine superiore delle sorgenti, non dai valori istantanei dei singoli stati. Questa dipendenza dalle statistiche di ordine superiore distingue V dalla funzione di energia delle reti di Hopfield (che dipende solo dai prodotti scalari degli stati) e richiede un approccio più sofisticato.

La costruzione si effettua in due passi. Nel primo passo, si definisce la funzione di coerenza istantanea $C(\Xi)$ come una misura scalare della coerenza delle rappresentazioni nello stato attuale Ξ :

$$C(\Xi) = \sum_{\{i < j\}} w_{ij} \cdot \tilde{C}(\xi_i, \xi_j) - \lambda \cdot \sum_i H(\xi_i)$$

dove $\tilde{C}(\xi_i, \xi_j) = I(\xi_i; \xi_j) / \sqrt{(H(\xi_i) \cdot H(\xi_j))}$ è la correlazione informazionale normalizzata tra le rappresentazioni ξ_i e ξ_j , w_{ij} sono pesi che riflettono la struttura delle connessioni anatomiche tra le aree, e il termine $-\lambda \cdot \sum_i H(\xi_i)$ penalizza la degenerazione delle rappresentazioni (rappresentazioni con bassa entropia sono poco informative). La funzione di energia è l'opposto della coerenza:

$$V(\Xi) = -C(\Xi) + V_{\text{ext}}(\Xi, E)$$

dove $V_{\text{ext}}(\Xi, E)$ è il termine che accoppia le rappresentazioni all'input ambientale E .

IV.5,3 — La struttura degli attrattori e la firma del contenuto

I minimi locali di V corrispondono agli attrattori della dinamica. La proprietà cruciale è che, in ciascun minimo Ξ^* , la firma informazionale $\Sigma(\Xi^*, \Theta)$ ha una struttura specifica: ha alta ridondanza (le rappresentazioni sono mutuamente coerenti e si corroborano a vicenda), moderata unicità (ciascuna rappresentazione porta qualche informazione non presente nelle altre, il che rende il sistema robusto alla perdita di singole componenti), e sinergia positiva (la struttura delle dipendenze tra le rappresentazioni cattura aspetti del contenuto che non sono presenti in nessuna rappresentazione individuale).

Si può dimostrare formalmente — sotto ipotesi ragionevoli sulla struttura di V — che i minimi di V corrispondono a massimi locali del coefficiente di coerenza $\kappa(\Xi, \Theta)$:

$$\nabla V(\Xi) = 0 \Leftrightarrow \partial \kappa(\Xi, \Theta) / \partial \Xi = 0$$

ovvero che gli stati stazionari della dinamica coincidono con i punti critici del coefficiente di coerenza. Questa equivalenza non è banale: richiede che la funzione di energia sia costruita in modo da rispettare la geometria dello spazio delle firme, il che è garantito dalla costruzione sopra descritta quando la metrica sulle rappresentazioni è compatibile con la metrica di Gromov-Hausdorff sulle firme.

IV.5,4 — Multistabilità e ambiguità percettiva: il caso del cubo di Necker formalizzato

La formalizzazione più precisa della multistabilità percettiva nel framework della funzione di energia si ottiene considerando uno stimolo visivo intrinsecamente ambiguo come il cubo di Necker. Il cubo di Necker è un disegno bidimensionale che è compatibile con due interpretazioni tridimensionali distinte: il cubo visto con la faccia frontale superiore destra (configurazione A) e il cubo visto con la faccia frontale inferiore sinistra (configurazione B). L'esperienza percettiva alterna tra le due interpretazioni con una periodicità tipicamente di 3-5 secondi.

Nel framework della funzione di energia, il sistema visivo ha una funzione V con due minimi locali, corrispondenti alle due interpretazioni. Le rappresentazioni nei due minimi hanno firme diverse:

$\Sigma_A(\Xi, \Theta)$: firma della configurazione A, con alta ridondanza tra le componenti che codificano la prospettiva «dall'alto a destra».

$\Sigma_B(\Xi, \Theta)$: firma della configurazione B, con alta ridondanza tra le componenti che codificano la prospettiva «dal basso a sinistra».

Tra i due minimi vi è una barriera energetica ΔV che separa i due bacini di attrazione. L'oscillazione percettiva corrisponde alle transizioni stocastiche tra i due minimi guidate dal termine di rumore $\eta(t)$ nell'equazione di evoluzione. La frequenza di alternanza è inversamente proporzionale all'altezza della barriera:

$$f_{alt} \approx (\sigma^2 / \Delta V) \cdot \nu$$

dove ν è un fattore di frequenza (analogo al fattore pre-esponenziale nell'equazione di Arrhenius per le reazioni chimiche). Questa formula produce la previsione quantitativa che la frequenza di alternanza aumenta con l'intensità del rumore σ^2 e diminuisce con l'altezza della barriera ΔV .

Il parametro ΔV dipende dalla struttura della funzione V , che a sua volta dipende dalla struttura di connessione w_i del sistema e dalle caratteristiche dello stimolo. Stimoli più ambigui (come un cubo di Necker sfocato o presentato per tempi brevi) producono barriere più basse e quindi alternanze più frequenti; stimoli meno ambigui (come un cubo con profondità enfatizzata attraverso ombre o gradienti) producono barriere più alte e quindi alternanze più rare.

La previsione formale sulla struttura dell'alternanza percettiva è quindi:

$$f_{alt}(\text{stimolo}) = (\sigma^2(\text{sistema}) / \Delta V(\text{stimolo}, \text{sistema})) \cdot v$$

dove $\sigma^2(\text{sistema})$ dipende dal livello di eccitabilità del sistema visivo (modulato dall'attenzione, dal carico cognitivo, dallo stato di affaticamento) e $\Delta V(\text{stimolo}, \text{sistema})$ dipende dall'interazione tra la struttura dello stimolo e la struttura di connessione del sistema. Questa formula è quantitativamente testabile attraverso la misura della frequenza di alternanza in funzione di manipolazioni sistematiche dello stimolo e dello stato del sistema.

IV.5,5 — Bistabilità nella cognizione sociale: un secondo caso formale

Un secondo caso di multistabilità cognitiva che si lascia trattare nel framework della funzione di energia è quello della percezione ambigua in cognizione sociale — la situazione in cui la stessa sequenza di comportamenti di un'altra persona può essere interpretata come espressione di emozioni diverse, come motivazioni diverse, o come caratteristiche di personalità diverse.

Consideriamo la percezione di un'espressione facciale borderline tra tristezza e disgusto — un'espressione che attiva in misura simile le rappresentazioni neurali della tristezza e del disgusto. Nel framework delle funzioni proprie, i sistemi neurali che codificano le espressioni emotive hanno funzioni V con due minimi corrispondenti a «tristezza» e «disgusto», con una barriera bassa per le espressioni borderline. In condizioni di percezione rapida (200 ms di esposizione), il sistema tende a convergere su uno dei due minimi in modo stabile e a non alternare. In condizioni di esposizione prolungata o di informazione ambientale ridotta, il sistema può alternare tra i due minimi, producendo l'esperienza di un'espressione «indefinita» o «mista».

La firma informazionale dei due minimi è diversa:

$\Sigma_{tristezza}(\Xi, \Theta)$: alta ridondanza tra le componenti che codificano la postura delle sopracciglia, l'angolo della bocca, e i movimenti degli occhi caratteristici della tristezza; componenti sinergiche con la rappresentazione della vicinanza interpersonale e dell'offerta di supporto.

$\Sigma_{disgusto}(\Xi, \Theta)$: alta ridondanza tra le componenti che codificano il grinzamento del naso, l'abbassamento delle labbra, e la tensione muscolare periorale del disgusto; componenti sinergiche con la rappresentazione della distanza interpersonale e dell'evitamento.

La previsione del modello è che il profilo di alternanza tra le due interpretazioni sia determinato dalla struttura delle barriere energetiche, che a sua volta dipende dalle caratteristiche fisiche dell'espressione (vicinanza alle caratteristiche prototipiche di tristezza e disgusto) e dalle

aspettative contestuali del percettore (che influenzano i pesi w_i della funzione V attraverso la modulazione top-down).

IV.5bis — La firma informazionale nel cervello: verso una teoria della misura neuroimaging

IV.5bis.1 — Dal modello computazionale alla misurazione empirica

La firma informazionale è un costrutto teorico-formale: è definita in termini di distribuzioni di probabilità e informazioni mutue che non sono direttamente osservabili nel cervello biologico. Per rendere la teoria empiricamente testabile, è necessario specificare un protocollo di traduzione dalla firma teorica alle misure neuroimaging accessibili. Questa sezione sviluppa tale protocollo in modo sufficientemente dettagliato da permettere la progettazione di esperimenti concreti.

Il protocollo si articola in quattro passi: la scelta delle variabili di sorgente e di obiettivo, la misura delle dipendenze statistiche tra le sorgenti e l'obiettivo, l'applicazione della decomposizione PID ai dati empirici, e l'interpretazione semantica della firma risultante.

IV.5bis.2 — La scelta delle variabili in esperimenti fMRI

Per applicare la PID ai dati di neuroimaging fMRI, le sorgenti $X_1, \dots, X_n$ sono le attività dei Regioni di Interesse (ROI) cerebrali selezionate sulla base di considerazioni anatomiche e funzionali, e l'obiettivo Y è la variabile dello stimolo o del compito che il soggetto sta elaborando. La scelta delle ROI richiede ipotesi teoriche su quali strutture neurali contribuiscono al contenuto in studio: questa scelta non è neutrale, perché ROI diverse producono firme diverse.

Per l'esempio del riconoscimento dei volti, le ROI rilevanti includono: V1/V2 (corteccia visiva primaria, processamento di basso livello), V4 (elaborazione del colore e della forma intermedia), FFA (fusiform face area, riconoscimento dell'identità del volto), STS (solco temporale superiore, elaborazione delle componenti dinamiche del volto), amigdala (valutazione emotiva), ippocampo (collocazione biografica), OFC/vmPFC (modello predittivo della persona). Queste sette ROI definiscono un sistema a $n=7$ sorgenti, con un numero di atomi PID nell'ordine di $2,4 \cdot 10^{12}$ ($D(7)$), che rende il calcolo esatto inaccessibile. Si utilizzerà dunque l'approssimazione gerarchica a coppie o la misura O-information per caratterizzare la struttura della firma in modo computazionalmente accessibile.

IV.5bis.3 — Stima delle informazioni mutue da serie temporali BOLD

Le serie temporali BOLD (Blood-Oxygen-Level-Dependent) ottenute da fMRI sono proxy delle attività neurali delle ROI, con due proprietà rilevanti per la stima della PID: sono serie temporali continue (non discrete) e sono rumorose (il segnale BOLD cattura soltanto una frazione dell'attività neurale sottostante). Per stimare le informazioni mutue tra le ROI e l'obiettivo Y , si utilizzano stimatori non-parametrici dell'informazione mutua basati sulla stima della densità di probabilità:

$$\hat{I}(X_i; Y) = \hat{H}(X_i) + \hat{H}(Y) - \hat{H}(X_i, Y)$$

dove $\hat{H}$ è un estimatore dell'entropia. La scelta dell'estimatore è critica: stimatori basati su istogrammi sono biased per campioni di dimensione finita; stimatori basati su kernel hanno problemi con la scelta della bandwidth; stimatori basati sul k-esimo vicino più prossimo (come il stimatore di Kraskov-Stögbauer-Grassberger) sono preferibili per le loro proprietà di correzione del bias.

Lo stimatore KSG (Kraskov-Stögbauer-Grassberger) dell'informazione mutua tra X e Y per un campione di N osservazioni è:

$$\hat{I}_{\text{KSG}}(X; Y) = \psi(k) + \psi(N) - \langle \psi(n_x(i) + 1) + \psi(n_y(i) + 1) \rangle_i$$

dove ψ è la funzione digamma, k è il numero di vicini considerati per ciascun punto, e $n_x(i)$ e $n_y(i)$ sono il numero di punti con distanza nell'asse X o Y inferiore alla distanza dal k -esimo vicino del punto i nello spazio congiunto. Questo stimatore ha bias $O(1/N)$ e varianza $O(1/N)$, il che lo rende accurato per campioni di dimensione $N \geq 100-200$, raggiungibili con sessioni fMRI di lunghezza moderata (30-60 minuti).

Con i valori $\hat{I}(X_i; Y)$ e $\hat{I}(X_i, X; Y)$ stimati per tutte le ROI e le loro coppie, si può calcolare l'O-information come misura surrogata della struttura della firma:

$$\Omega(X_1, \dots, X; Y) = (n-2)\hat{I}(X_1, \dots, X; Y) + \sum_i \hat{I}(X_i; Y) - \sum_i \hat{I}(X_1, \dots, \hat{X}_i, \dots, X; Y)$$

e le sinergie di coppia $\tilde{S}(X_i, X; Y)$ come misure della struttura locale della firma.

IV.5bis.4 — Interpretazione della firma neuroimaging e validazione

La firma neuroimaging stimata attraverso il protocollo sopra descritto è una caratterizzazione parziale della firma informazionale teorica: cattura le componenti di ridondanza e sinergia tra le ROI selezionate rispetto all'obiettivo scelto, ma non l'intera struttura del reticolo degli atomi PID.

La valutazione della qualità di questa caratterizzazione parziale richiede criteri di validazione che si articolano a più livelli.

Il primo livello è la validazione interna: si verifica che la firma sia stabile attraverso sessioni diverse dello stesso soggetto, che sia robusta rispetto alla scelta del parametro k nell'estimatore KSG, e che le sue componenti riflettano le ipotesi anatomiche e funzionali sulla connettività tra le ROI. Una firma con alta ridondanza tra ROI strutturalmente connesse (come FFA e STS, che hanno connessioni dirette attraverso il fascicolo longitudinale inferiore) e bassa ridondanza tra ROI strutturalmente distanti (come V1 e OFC, che non hanno connessioni dirette) è coerente con le ipotesi anatomiche ed è un segnale di validità interna.

Il secondo livello è la validazione predittiva: si verifica che la firma predica comportamenti futuri del soggetto in compiti che richiedono l'elaborazione del contenuto in studio. Un soggetto con alta sinergia tra FFA e amigdala per i volti dovrebbe mostrare migliori prestazioni in compiti che richiedono l'integrazione dell'identità e della valenza emotiva dei volti (come il riconoscimento della relazione emotiva tra due persone in una scena sociale). Se la sinergia misura informazione che emerge solo dall'analisi congiunta di FFA e amigdala, allora questa sinergia dovrebbe essere predittiva dei comportamenti che richiedono quell'integrazione.

Il terzo livello è la validazione across-subjects: si verifica che le firme di soggetti con simile expertise (ad esempio, psicologi clinici esperti nel riconoscimento delle espressioni emotive) abbiano distanze di Gromov-Hausdorff più piccole tra loro rispetto alle distanze tra soggetti con expertise diversa. Se la firma misura il contenuto categoriale, soggetti con lo stesso tipo di expertise dovrebbero avere firme più simili rispetto a soggetti con expertise diverse.

Questi tre livelli di validazione definiscono un programma di ricerca empirico concreto che trasforma la teoria formale della firma informazionale in un programma di neuroimaging computazionale — un programma che è già in parte realizzabile con le tecnologie e i metodi statistici attualmente disponibili, e che diventerà più accessibile con lo sviluppo di metodi di stima della PID ad alta dimensionalità.

IV.5 — Coerenza come attrattore: la dinamica delle rappresentazioni ridondanti

La firma informazionale definita nella sezione precedente è una misura statica: descrive la struttura informatica del sistema in un dato momento, per un dato contenuto. Per costruire una teoria dinamica — che renda conto di come il sistema mantiene la coerenza attraverso il tempo e di come la aggiorna in risposta a nuovi input — è necessario modellare la dinamica delle rappresentazioni ridondanti come un processo che tende verso stati di coerenza elevata.

Si consideri un sistema cognitivo con n rappresentazioni ridondanti $\xi_1(t), \dots, \xi_n(t)$ che evolvono nel tempo. Ciascuna rappresentazione riceve input diretto dall'ambiente $E(t)$ e input dalle altre rappresentazioni attraverso connessioni sia feed-forward sia feed-back:

$$d\xi_i/dt = f_i(\xi_i, E) + \sum_{j \neq i} w_{ij} \cdot g_{ij}(\xi_i, \xi_j) + \eta_i(t)$$

dove f_i è la risposta di ξ_i all'input ambientale E , w_{ij} sono i pesi di accoppiamento tra le rappresentazioni ξ_i e ξ_j , g_{ij} è la funzione di interazione locale tra le due rappresentazioni, e $\eta_i(t)$ è il termine di rumore gaussiano con matrice di covarianza Q_i . Il termine di accoppiamento $\sum w_{ij} \cdot g_{ij}(\xi_i, \xi_j)$ cattura l'influenza reciproca tra rappresentazioni ridondanti: le rappresentazioni non sono indipendenti ma si co-vincolano.

Per caratterizzare la struttura attrattiva di questo sistema dinamico, si introduce una funzione di Lyapunov $V(\Xi)$ che ha minimi locali corrispondenti agli stati di massima coerenza:

$$V(\Xi) = -TC(\Xi | \Theta) + \lambda \cdot \sum_i \|\xi_i - E\|^2$$

dove $TC(\Xi | \Theta) = \sum_i H(\xi_i | \Theta) - H(\xi_1, \dots, \xi_n | \Theta)$ è la co-informazione totale condizionale — la quantità che misura la dipendenza mutua tra le rappresentazioni condizionatamente al contenuto — e il secondo termine penalizza la distanza dall'input ambientale E con peso $\lambda > 0$. Il sistema evolve verso minimi locali di V , che corrispondono a stati in cui la co-informazione condizionale è massimizzata compatibilmente con il vincolo di fedeltà all'input.

Questo sistema dinamico ha tre proprietà rilevanti per la teoria del contenuto.

La prima è la multistabilità: per certi valori dei parametri di accoppiamento w_{ij} e del peso λ , la funzione V ha più minimi locali, corrispondenti a stati di coerenza diversi compatibili con lo stesso input ambientale. La multistabilità è il correlato formale dell'ambiguità percettiva: uno stimolo ambiguo è compatibile con più stati di coerenza del sistema rappresentazionale, e la percezione alterna tra essi secondo dinamiche stocastiche. Il tempo di transizione tra i minimi — il tempo di Kramers — è inversamente proporzionale alla differenza di profondità dei minimi nella funzione di energia, il che spiega perché alcune ambiguità percettive si risolvono rapidamente mentre altre persistono a lungo.

La seconda proprietà è l'isteresi: il sistema tende a rimanere nel bacino di attrazione del minimo in cui si trova, resistendo a piccole perturbazioni dell'input. Formalmente, un piccolo cambiamento δE nell'input non sposta il sistema sul minimo di V più vicino a $E + \delta E$, ma lo mantiene nel bacino di attrazione del minimo precedente fino a che la perturbazione supera la barriera energetica tra i bacini. Questo

corrisponde alla stabilità percettiva in condizioni di input variabile — la costanza percettiva — e alla resistenza alla modifica delle credenze in presenza di evidenza debolmente contraria.

La terza proprietà è la capacità di completamento: se l'input è parziale o rumoroso — se alcune componenti dell'input ambientale E sono mancanti o corrotte — il sistema tende verso il minimo di V più vicino allo stato attuale, completando le rappresentazioni mancanti sulla base della struttura di coerenza già stabilita. Questo corrisponde al completamento percettivo, all'inferenza abduttiva e alla memoria ricostruttiva.

La tesi della ridondanza costitutiva, in termini dinamici, è che il contenuto di un sistema rappresentazionale sia identificabile con i minimi stabili della funzione V — gli attrattori del processo di aggiornamento. Formalmente, il contenuto Θ associato a un attrattore Ξ^* è definito come:

$$\Theta(\Xi^*) \equiv \operatorname{argmax}_{\{\Theta'\}} \kappa(\Xi^*, \Theta')$$

cioè il contenuto che massimizza il coefficiente di coerenza della firma informazionale dell'attrattore. Questa definizione implica che il contenuto di un sistema rappresentazionale non è fissato dall'esterno — non è una proprietà che le rappresentazioni hanno indipendentemente dalla loro struttura di coerenza — ma emerge dalla struttura interna dell'attrattore. Due sistemi con attrattori diversi hanno contenuti diversi, anche se le loro rappresentazioni individuali sembrano simili; due sistemi con attrattori simili hanno contenuti simili, anche se le loro rappresentazioni individuali sono codificate in modi superficialmente molto diversi.

IV.6 — Misura della coerenza per la ridondanza gerarchica

La struttura formale della PID si applica naturalmente alla ridondanza orizzontale — la coesistenza di rappresentazioni allo stesso livello della gerarchia cognitiva. Per la ridondanza gerarchica — la coesistenza di rappresentazioni a livelli diversi della gerarchia predittiva — è necessario un apparato leggermente diverso che tenga conto della struttura verticale delle relazioni tra livelli.

Nel framework della codifica predittiva gerarchica, il modello generativo specifica una distribuzione congiunta su osservazioni e stati latenti a tutti i livelli:

$$P(o, s^{(1)}, \dots, s^{(L)}) = P(o \mid s^{(1)}) \cdot \prod_{l=1}^{L-1} P(s^{(l+1)} \mid s^{(l)}) \cdot P(s^{(L)})$$

Ciascun livello $s^{(l)}$ è una rappresentazione del contenuto al livello di astrazione l . Le rappresentazioni a livelli diversi sono statisticamente dipendenti: condizionatamente alla distribuzione congiunta del modello generativo, $s^{(l)}$ e $s^{(l')}$ sono dipendenti per ogni $l \neq l'$.

Per misurare la struttura di ridondanza gerarchica, si considera l'informazione mutua tra il livello l e il contenuto Θ :

$$I(s^{(l)}; \Theta) \text{ per ogni } l = 1, \dots, L$$

e la struttura delle dipendenze tra livelli condizionata al contenuto:

$$I(s^{(l)}; s^{(l')} \mid \Theta) \text{ per } l \neq l'$$

Questa quantità misura quanto i livelli l e l' condividono informazione sulla struttura causale del mondo che va oltre l'informazione già catturata dal contenuto Θ — una misura della ridondanza gerarchica tra livelli. Sotto l'ipotesi markoviana del modello generativo, si ha la fattorizzazione:

$$I(s^{(l)}; s^{(l')} \mid \Theta) = I(s^{(l)}; s^{(l'-1)} \mid s^{(l'+1)}, \Theta) + \dots$$

che mostra come la ridondanza gerarchica si struttura attraverso le catene di dipendenza del modello generativo. La decomposizione PID applicata ai livelli della gerarchia produce una struttura reticolare degli atomi informativi gerarchici che generalizza la struttura piatta del caso non gerarchico.

Un risultato formale di rilievo riguarda la relazione tra la struttura di ridondanza gerarchica e il gradiente dell'energia libera variazionale. Nelle equazioni di aggiornamento variazionale standard del modello di Friston:

$$\mu^{(l)} = (\partial g^{(l)} / \partial \mu^{(l)})^T \xi^{(l)} - \xi^{(l+1)}$$

il termine $\xi^{(l)} = \Pi^{(l)} \varepsilon^{(l)}$ è l'errore di previsione pesato per la precisione al livello l . La struttura di ridondanza gerarchica — misurata dalla PID applicata ai livelli del modello — corrisponde alla struttura delle correlazioni tra gli errori di previsione a livelli diversi. Questa corrispondenza permette di derivare previsioni empiriche sulla struttura degli errori di previsione dai dati neurobiologici.

IV.7 — L'identità del contenuto come invariante della firma informativa

Arriviamo alla definizione formale dell'identità del contenuto che la tesi della ridondanza costitutiva offre. Due stati cognitivi s_1 e s_2 — che possono appartenere allo stesso sistema in momenti diversi, o a sistemi cognitivi diversi — hanno lo stesso contenuto rispetto all'obiettivo Θ se

e solo se le loro firme informazionali sono isomorfe come strutture reticolari pesate:

$$\text{Cont}(s_1) = \text{Cont}(s_2) \text{ iff } \Sigma(\Xi(s_1), \Theta) \cong \Sigma(\Xi(s_2), \Theta)$$

dove l'isomorfismo è un isomorfismo di reticoli pesati: una biiezione $\varphi: \mathcal{A}_1 \rightarrow \mathcal{A}_2$ tra i reticoli degli atomi dei due sistemi che preserva l'ordinamento parziale e i valori dei pesi degli atomi:

$$\varphi(\alpha) \leq \varphi(\beta) \text{ iff } \alpha \leq \beta, \text{ e } \mathcal{R}(\varphi(\alpha); \Theta)_2 = \mathcal{R}(\alpha; \Theta)_1 \text{ per ogni } \alpha \in \mathcal{A}_1$$

Questa definizione di identità del contenuto ha diverse proprietà desiderabili. In primo luogo, è invariante rispetto alle implementazioni: due sistemi che implementano lo stesso contenuto attraverso substrati fisici completamente diversi — un cervello umano e un sistema artificiale, o due cervelli con architetture di connettività diverse — hanno lo stesso contenuto se e solo se le loro firme informazionali sono isomorfe, indipendentemente dai dettagli dell'implementazione. In secondo luogo, è graduale: due sistemi possono avere firme «quasi isomorfe» nel senso della distanza di Gromov-Hausdorff che si definirà nel prossimo capitolo, e in tal caso si può dire che hanno contenuti «simili» in un senso preciso e misurabile. In terzo luogo, è falsificabile: si possono progettare esperimenti neurobiologici o comportamentali che testano se due stati cognitivi hanno la stessa firma informazionale rispetto a un dato obiettivo.

La definizione ha anche un costo: richiede la scelta dell'obiettivo Θ rispetto al quale la firma è definita. L'obiettivo non è parte della struttura cognitiva interna ma è una variabile del mondo esterno; la sua scelta dipende quindi da una teoria di cosa conta come «il contenuto» che lo stato dovrebbe rappresentare. Questa dipendenza dall'obiettivo riflette il fatto che il contenuto è sempre contenuto-di-qualcosa: la firma informazionale di uno stato cognitivo rispetto all'obiettivo Θ_1 può essere completamente diversa dalla sua firma rispetto all'obiettivo Θ_2 , e la scelta dell'obiettivo rilevante dipende dal contesto teorico e pragmatico in cui la domanda sul contenuto è posta. Questo non è un difetto della teoria; è una rispecchiatura della struttura prospettiva del contenuto mentale, che è sempre contenuto-per-un-soggetto-in-un-contesto.

Capitolo V — La geometria dei contenuti: distanza, variazione e oggettività della firma informazionale

V.1 — La necessità di una metrica sullo spazio dei contenuti

La firma informazionale definita nel Capitolo IV è un oggetto matematico che vive in uno spazio di firme — lo spazio di tutte le possibili triple $(\mathcal{R}, \kappa, \sigma)$ compatibili con le definizioni date. Per usare questo oggetto come base di una teoria del contenuto, è necessario dotare lo spazio delle firme di una struttura metrica che permetta di misurare la distanza tra firme diverse e, di conseguenza, la distanza tra contenuti diversi. La necessità di questa struttura metrica non è puramente formale; riflette l'intuizione sostanziale che i contenuti variano gradualmente e che la variazione del contenuto è misurabile.

L'intuizione può essere illustrata con un esempio. Se un soggetto umano impara che un oggetto che credeva fosse un uccello è in realtà un pipistrello — un mammifero — la sua rappresentazione dell'oggetto cambia, e questo cambiamento è un cambiamento di contenuto. Ma non è un cambiamento di tutto il contenuto: molte delle rappresentazioni ridondanti dell'oggetto — la sua forma, il suo comportamento, la sua collocazione ecologica — rimangono invariate, mentre le rappresentazioni che codificano la sua appartenenza tassonomica cambiano radicalmente. Una teoria del contenuto che non possa misurare la distanza tra i contenuti prima e dopo l'apprendimento non può caratterizzare questo tipo di cambiamento in modo informativo.

V.2 — Metriche sullo spazio delle distribuzioni di probabilità: geometria dell'informazione

Prima di definire la distanza tra firme informazionali, è necessario definire la distanza tra distribuzioni di probabilità, che è la materia prima degli atomi PID. La scelta della metrica è una scelta teorica: diverse metriche catturano diverse intuizioni su cosa significhi che due distribuzioni siano «simili» o «distanti».

La geometria dell'informazione nel senso di Amari fornisce la struttura più appropriata per il presente scopo. Lo spazio delle distribuzioni di probabilità su un alfabeto finito Ω — che denotiamo $\Delta(\Omega)$ — è una varietà differenziabile di dimensione $|\Omega| - 1$. Su questa varietà si può definire una metrica riemanniana derivata dalla matrice di Fisher:

$$g_{ij}(\mathbf{p}) = \sum_{\omega \in \Omega} p(\omega) \cdot (\partial \log p(\omega) / \partial \theta_i) \cdot (\partial \log p(\omega) / \partial \theta_j)$$

dove $\theta = (\theta_1, \dots, \theta_{|\Omega|-1})$ sono le coordinate locali. La distanza geodetica sul rimanifold riemanniano $(\Delta(\Omega), g)$ è la distanza di Fisher-Rao:

$$d_{\text{FR}}(p, q) = 2 \cdot \arccos \left(\sum_{\omega \in \Omega} \sqrt{p(\omega) \cdot q(\omega)} \right)$$

La distanza di Fisher-Rao ha la proprietà di invarianza sotto trasformazioni sufficienti: se T è una statistica sufficiente per il parametro θ , allora $d_{\text{FR}}(p_{\theta_1}, p_{\theta_2}) = d_{\text{FR}}(p_{\{T, \theta_1\}}, p_{\{T, \theta_2\}})$. Questa proprietà è desiderabile per una metrica sullo spazio dei contenuti perché garantisce che la distanza tra contenuti non dipende dalla scelta della rappresentazione — che è precisamente la proprietà di invarianza rispetto all'implementazione che si richiede.

Un'alternativa alla distanza di Fisher-Rao, che ha proprietà computazionali più maneggevoli, è la divergenza di Jensen-Shannon simmetrizzata:

$$\text{JSD}(p \parallel q) = (1/2) \cdot \text{KL}(p \parallel m) + (1/2) \cdot \text{KL}(q \parallel m)$$

dove $m = (p + q)/2$ è la distribuzione media e KL è la divergenza di Kullback-Leibler $\text{KL}(p \parallel q) = \sum_{\omega} p(\omega) \log(p(\omega)/q(\omega))$. La divergenza JSD è simmetrica, sempre finita (anche quando KL non lo è) e ha un significato informativo preciso: misura la quantità di informazione necessaria per distinguere tra p e q quando il tipo di distribuzione è sconosciuto a priori. La radice quadrata della JSD è una vera distanza metrica sullo spazio delle distribuzioni, nota come distanza di Jensen-Shannon.

V.3 — La distanza di Gromov-Hausdorff tra firme informazionali

Con una metrica sullo spazio delle distribuzioni di probabilità disponibile, si può ora definire la distanza tra firme informazionali. La difficoltà principale è che due firme possono avere dimensioni diverse — corrispondenti a sistemi con un numero diverso di rappresentazioni ridondanti — e la comparazione diretta tra vettori di dimensioni diverse richiederebbe una mappa tra gli spazi degli atomi che potrebbe non essere ovvia.

La distanza di Gromov-Hausdorff fornisce uno strumento matematico appropriato per questo tipo di comparazione. La distanza di Hausdorff tra due sottoinsiemi compatti A e B di uno spazio metrico (M, d) è:

$$d_H(A, B) = \max \left(\sup_{a \in A} d(a, B), \sup_{b \in B} d(b, A) \right)$$

dove $d(a, B) = \inf_{b \in B} d(a, b)$ è la distanza del punto a dall'insieme B . La distanza di Gromov-Hausdorff generalizza questa nozione a spazi metrici che non vivono in uno spazio ambiente comune:

$$d_{\text{GH}}(X, Y) = \inf_{\{Z, f, g\}} d_H(f(X), g(Y))$$

dove l'infimo è preso su tutti gli spazi metrici Z e tutte le isometrie parziali $f: X \rightarrow Z$ e $g: Y \rightarrow Z$. La distanza di Gromov-Hausdorff misura quanto un spazio metrico può essere «reso simile» a un altro attraverso una mappa che rispetti le distanze interne.

Per le firme informazionali, la distanza di Gromov-Hausdorff si calcola come segue. Siano $\Sigma_A = (\mathcal{R}_A, \kappa_A, \sigma_A)$ e $\Sigma_B = (\mathcal{R}_B, \kappa_B, \sigma_B)$ due firme corrispondenti a sistemi con n_A e n_B rappresentazioni ridondanti, rispettivamente. I reticoli degli atomi $\mathcal{R}_A$ e $\mathcal{R}_B$ sono spazi metrici con la metrica:

$$d_{\text{atom}}(\alpha, \beta) = |\mathcal{R}(\alpha; \Theta)_A - \mathcal{R}(\beta; \Theta)_B| \text{ per } \alpha \in \mathcal{A}_{\{n_A\}}, \beta \in \mathcal{A}_{\{n_B\}}$$

La distanza tra le firme è:

$$d_{\text{GH}}(\Sigma_A, \Sigma_B) = \inf_{\varphi} \{ \sup_{\alpha \in \mathcal{A}_{\{n_A\}}} |\mathcal{R}(\alpha; \Theta)_A - \mathcal{R}(\varphi(\alpha); \Theta)_B| \}$$

dove l'infimo è preso su tutte le biiezioni preservanti l'ordine $\varphi: \mathcal{A}_{\{n_A\}} \rightarrow \mathcal{A}_{\{n_B\}}$ — gli isomorfismi di reticoli — quando tali biiezioni esistono. Se i reticoli non sono isomorfi come reticoli astratti, la distanza è fissata a un valore massimo di default (che corrisponde all'intuizione che contenuti con strutture reticolari completamente diverse sono massimalmente distanti).

Questa metrica ha le proprietà che si richiedono per uno spazio dei contenuti: è simmetrica, soddisfa la disuguaglianza triangolare, ed è zero se e solo se le firme sono identiche (cioè se e solo se le strutture reticolari pesate sono isomorfe). Dotato di questa metrica, lo spazio delle firme informazionali diventa uno spazio metrico su cui si possono definire nozioni di vicinanza, traiettoria e convergenza — e il cambiamento di contenuto nel tempo può essere caratterizzato come una traiettoria in questo spazio.

V.4 — Traiettorie nel spazio dei contenuti e cambiamento concettuale

La possibilità di misurare la distanza tra firme informazionali permette di caratterizzare il cambiamento di contenuto come una traiettoria nello spazio delle firme. Sia $\Sigma(t) = \Sigma(\Xi(t), \Theta)$ la firma informazionale del sistema cognitivo al tempo t rispetto al contenuto Θ . Il cambiamento di contenuto nel periodo $[t_0, t_1]$ è la traiettoria $\{\Sigma(t) : t \in [t_0, t_1]\}$ nello spazio delle firme.

Una proprietà importante di questa traiettoria è la sua velocità: la derivata della firma rispetto al tempo, che misura la rapidità con cui il contenuto si modifica. La velocità di cambiamento è:

$$|d\Sigma/dt| = \lim_{\epsilon \rightarrow 0} \{ d_{\text{GH}}(\Sigma(t + \epsilon), \Sigma(t)) / \epsilon \}$$

I periodi di apprendimento rapido corrispondono a velocità elevata; i periodi di consolidazione corrispondono a velocità ridotta. Il cambiamento concettuale radicale — la rivoluzione scientifica nel senso di Kuhn, o la riorganizzazione delle credenze in seguito a una forte evidenza controintuitiva — corrisponde a un'accelerazione brusca della traiettoria seguita da una stabilizzazione su una nuova firma di attrazione.

La continuità della traiettoria — o la sua discontinuità — ha un significato per l'identità del contenuto nel tempo. Si può definire la misura di continuità identitaria di una traiettoria come:

$$C([\Sigma_{t_0}, \Sigma_{t_1}]) = 1 - \sup_{t \in [t_0, t_1]} d_{GH}(\Sigma(t), \Sigma(t_0)) / D_{max}$$

dove D_{max} è la distanza massima possibile nello spazio delle firme. Questa misura è 1 quando la traiettoria è completamente stazionaria (il contenuto non cambia), e tende a 0 quando la traiettoria attraversa regioni dello spazio delle firme molto lontane dalla firma iniziale. Il valore di C caratterizza in modo continuo il grado in cui il contenuto al tempo t_1 è «lo stesso contenuto» di quello al tempo t_0 .

V.5 — L'oggettività del contenuto e il grado di intersoggettività

Una delle domande più importanti che una teoria del contenuto deve affrontare è quella dell'oggettività: in che senso vi è un contenuto «oggettivo» che diversi sistemi cognitivi possono condividere? La risposta che il quadro della ridondanza costitutiva offre non è né il realismo ingenuo — che afferma l'esistenza di un contenuto unico e determinato indipendentemente dai sistemi cognitivi che lo rappresentano — né il relativismo radicale — che afferma che ogni sistema cognitivo ha il proprio contenuto, incomparabile con quello degli altri.

Il grado di oggettività di un contenuto Θ rispetto a una famiglia di N sistemi cognitivi $\{A_1, \dots, A_N\}$ — che possono essere individui diversi, o lo stesso individuo in tempi diversi, o sistemi cognitivi artificiali — è definito come:

$$\Omega(\Theta) = 1 - (1 / N(N-1)) \cdot \sum_{i \neq j} d_{GH}(\Sigma_{A_i}(\Theta), \Sigma_{A_j}(\Theta)) / \max_{i,j} d_{GH}(\Sigma_{A_i}(\Theta), \Sigma_{A_j}(\Theta))$$

dove $\Sigma_{A_i}(\Theta)$ è la firma informazionale del sistema A_i rispetto al contenuto Θ . La misura Ω è 1 quando tutte le firme sono identiche — quando tutti i sistemi rappresentano Θ «nello stesso modo» — e tende a 0 quando le firme sono massimamente diverse. L'oggettività è dunque una proprietà graduale: non si ha oggettività piena o oggettività nulla, ma un continuum di gradi di allineamento tra le firme dei diversi sistemi.

Questa caratterizzazione dell'oggettività ha due implicazioni importanti. La prima è che l'oggettività dipende dal dominio di riferimento: il

contenuto «acqua è H_2O » ha un grado di oggettività elevato tra tutti i chimici che hanno appreso la struttura molecolare dell'acqua, un grado inferiore tra tutte le persone colte, e un grado ancora inferiore tra tutta la popolazione umana. Non vi è un grado di oggettività «assoluto», indipendente dalla scelta della famiglia di sistemi rispetto a cui si misura l'allineamento. La seconda implicazione è che l'oggettività è empiricamente misurabile: si può testare se diversi individui hanno firme informazionali allineate rispetto a un dato contenuto attraverso paradigmi sperimentali che misurano le strutture di dipendenza tra rappresentazioni modali e aspettuali diverse.

V.6 — Topologia persistente e invarianti topologici della firma informazionale

V.6,1 — Il problema della struttura globale

Le sezioni precedenti di questo capitolo hanno fornito gli strumenti per misurare distanze tra firme informazionali e per tracciare traiettorie nello spazio dei contenuti. Tuttavia, la distanza di Gromov-Hausdorff tra firme, per quanto sofisticata, cattura essenzialmente proprietà locali della struttura reticolare: misura quanto due firme differiscano negli atomi corrispondenti e nelle loro relazioni immediate. Esistono proprietà della struttura informazionale di un sistema che sono genuinamente globali nel senso topologico — proprietà che non emergono dall'analisi degli atomi locali e delle loro relazioni prossimali, ma dalla struttura dell'intero reticolo considerato come spazio geometrico. Queste proprietà globali includono la presenza di cluster distinti di atomi (componenti connesse), la presenza di cicli informativi (percorsi chiusi attraverso il reticolo degli atomi che non sono riducibili a cammini più semplici), la presenza di cavità tridimensionali (strutture vuote circondate da atomi che le delimitano ma non le colmano), e strutture di ordine superiore di dimensione k corrispondenti a correlazioni informatiche irriducibili tra $k+1$ atomi simultaneamente.

Queste strutture globali hanno un'interpretazione diretta in termini di contenuto. I cluster di atomi corrispondono a componenti semanticamente coese del contenuto — sottostrutture della firma che si organizzano in unità quasi-indipendenti. I cicli informativi corrispondono a relazioni circolari di dipendenza tra aspetti del contenuto: strutture in cui A è informativo su B , B è informativo su C , e C è informativo su A , senza che questa circolarità sia riducibile a dipendenze dirette. Le cavità tridimensionali corrispondono a «buchi» nel contenuto — regioni dello spazio informazionale che sono delimitate da atomi ma non riempite da altri atomi — e hanno un'interpretazione come zone di incertezza strutturata, regioni del dominio di contenuto che il sistema riconosce come esistenti (perché gli atomi che le delimitano sono presenti) ma non rappresenta direttamente.

La topologia algebrica computazionale, e specificamente la tecnica dell'omologia persistente, fornisce gli strumenti per estrarre, classificare e quantificare queste strutture globali da spazi metrici ad alta dimensionalità come il reticolo degli atomi PID della firma informazionale. L'omologia persistente è stata sviluppata nel contesto dell'analisi dei dati topologici (TDA, Topological Data Analysis) da Edelsbrunner, Letscher e Zomorodian agli inizi del decennio 2000, e ha trovato applicazioni in biologia computazionale, analisi di immagini mediche, analisi di reti complesse e, più recentemente, nella neuroscienza computazionale. Il suo impiego per l'analisi della struttura topologica delle firme informazionali è una proposta nuova che si sviluppa nella presente sezione, con derivazioni formali esplicite e applicazioni a casi concreti costruiti.

V.6,2 — Costruzione del complesso simpliciale filtrato sul reticolo degli atomi

Il punto di partenza è il reticolo degli atomi $\mathcal{R}(\Xi, \Theta)$ della firma informazionale di un sistema Ξ rispetto all'obiettivo Θ . Come si è definito nel Capitolo IV, questo reticolo è un insieme parzialmente ordinato i cui elementi — gli atomi — corrispondono a contribuzioni informazionali irriducibili di diversi sottoinsiemi delle sorgenti. Ogni atomo α ha un peso $\omega(\alpha) = \mathcal{R}(\alpha; \Theta)$ che misura la sua contribuzione all'informazione mutua totale. Per rendere questo reticolo un oggetto geometrico su cui applicare la topologia persistente, è necessario dotarlo di una struttura metrica e costruire su di esso una famiglia di complessi simpliciali che approssimino la sua struttura a diverse scale di risoluzione.

La metrica sul reticolo degli atomi è definita nel modo seguente. Due atomi $\alpha, \beta \in \mathcal{A}$ sono vicini se le loro distribuzioni congiunte condizionali sono simili, dove la distribuzione condizionale dell'atomo α rispetto all'obiettivo Θ è la distribuzione $P(\Theta | \alpha \text{ si realizza})$ che cattura il modo in cui l'atomo contribuisce all'informazione sull'obiettivo. La distanza tra atomi è:

$$d(\alpha, \beta) = d_{\text{JS}}(P_{\alpha}(\Theta), P_{\beta}(\Theta))$$

dove d_{JS} è la distanza di Jensen-Shannon tra le distribuzioni condizionali associate a ciascun atomo, definita come:

$$d_{\text{JS}}(P_{\alpha}, P_{\beta}) = \sqrt{\text{JSD}(P_{\alpha} \parallel P_{\beta})}$$

con $\text{JSD}(P_{\alpha} \parallel P_{\beta}) = (1/2) \text{KL}(P_{\alpha} \parallel M) + (1/2) \text{KL}(P_{\beta} \parallel M)$ e $M = (P_{\alpha} + P_{\beta})/2$.

Con questa metrica, il reticolo degli atomi diventa uno spazio metrico finito $(\mathcal{A}, d)$. La costruzione del complesso simpliciale filtrato procede attraverso la costruzione del complesso di Vietoris-Rips (o, alternativamente, del complesso di Čech, che ha migliori proprietà teoriche ma è computazionalmente più oneroso).

Il complesso di Vietoris-Rips $VR(\mathcal{A}, \epsilon)$ a scala ϵ è definito come segue:

$$VR(\mathcal{A}, \epsilon) = \{ \sigma \subseteq \mathcal{A} : d(\alpha, \beta) \leq \epsilon \text{ per ogni } \alpha, \beta \in \sigma \}$$

In altri termini, un sottoinsieme σ di atomi forma un simpleso nel complesso $VR(\mathcal{A}, \epsilon)$ se e soltanto se ogni coppia di atomi in σ ha distanza al più ϵ . Al crescere di ϵ da 0 a ∞ , il complesso $VR(\mathcal{A}, \epsilon)$ cresce in modo monotono (ogni simpleso presente a scala ϵ è presente anche a scala $\epsilon' > \epsilon$), producendo una sequenza annidata di complessi:

$$\emptyset = VR(\mathcal{A}, 0) \subseteq VR(\mathcal{A}, \epsilon_1) \subseteq VR(\mathcal{A}, \epsilon_2) \subseteq \dots \subseteq VR(\mathcal{A}, \infty) = \Delta(\mathcal{A})$$

dove $\Delta(\mathcal{A})$ è il simpleso pieno su tutti gli atomi. Questa sequenza annidata è la filtrazione Vietoris-Rips di $(\mathcal{A}, d)$, e l'omologia persistente studia come la struttura topologica di $VR(\mathcal{A}, \epsilon)$ varia al variare di ϵ .

Il complesso di Čech $\check{Cech}(\mathcal{A}, \epsilon)$ ha una costruzione alternativa basata sulle sfere: σ è un simpleso in $\check{Cech}(\mathcal{A}, \epsilon)$ se e solo se le sfere di raggio ϵ centrate negli atomi di σ hanno un'intersezione comune non vuota. Per il Lemma del Nervo, il complesso di Čech è omotopicamente equivalente all'unione delle sfere chiuse di raggio ϵ centrate negli atomi. La differenza tra VR e Čech è dunque nella precisione della approssimazione geometrica: VR è più facile da calcolare ma può sovrastimare la struttura topologica, mentre Čech cattura in modo più preciso la geometria dello spazio metrico sottostante.

V.6,3 — L'omologia simpliciale e i numeri di Betti

L'omologia di un complesso simpliciale K è uno strumento algebrico che misura in modo quantitativo le strutture topologiche del complesso a diversi livelli dimensionali. Il meccanismo formale si basa sui gruppi di omologia $H_k(K) = \ker \partial_k / \text{im } \partial_{k+1}$, dove $\partial_k : C_k(K) \rightarrow C_{k-1}(K)$ è l'operatore bordo che mappa ogni k -simpleso nella combinazione alternata dei suoi $(k-1)$ -facciali.

Procediamo con la costruzione esplicita. Per un complesso K con insieme di vertici $\mathcal{A}$, i moduli delle catene sono definiti nel modo seguente. Il modulo delle 0-catene $C_0(K)$ è il modulo libero generato dai vertici: $C_0(K) = \mathbb{Z}^{\mathcal{A}}$. Il modulo delle 1-catene $C_1(K)$ è il modulo libero generato dagli spigoli orientati (coppie ordinate di vertici adiacenti): $C_1(K) = \mathbb{Z}^{\{E(K)\}}$. Il modulo delle k -catene $C_k(K)$ è il modulo libero generato dai k -simplessi orientati: $C_k(K) = \mathbb{Z}^{\{K_k\}}$ dove K_k è l'insieme dei k -simplessi di K .

L'operatore bordo $\partial_1 : C_1(K) \rightarrow C_0(K)$ mappa ogni spigolo orientato (α, β) nella differenza $\partial_1(\alpha, \beta) = \beta - \alpha$. L'operatore bordo $\partial_2 : C_2(K) \rightarrow C_1(K)$ mappa ogni triangolo orientato (α, β, γ) nella combinazione alternata dei suoi tre lati: $\partial_2(\alpha, \beta, \gamma) = (\beta, \gamma) - (\alpha, \gamma) + (\alpha, \beta)$. In generale,

l'operatore bordo ∂_k mappa un k -simpleso orientato $(v_0, v_1, \dots, v_k)$ nella combinazione alternata dei suoi $(k-1)$ -facciali:

$$\partial_k(v_0, \dots, v_k) = \sum_i (-1)^i (v_0, \dots, \hat{v}_i, \dots, v_k)$$

dove $\hat{v}_i$ indica che il vertice v_i è omissso. La proprietà fondamentale dell'operatore bordo è $\partial_k \circ \partial_{k+1} = 0$ per ogni k , che garantisce $\text{im } \partial_{k+1} \subseteq \ker \partial_k$ e rende ben definita l'omologia.

Il k -esimo gruppo di omologia è:

$$H_k(K; Z) = \ker \partial_k / \text{im } \partial_{k+1}$$

Il k -esimo numero di Betti β_k è il rango del gruppo di omologia (considerato come Z -modulo libero, ignorando i gruppi di torsione che sono infrequenti nelle applicazioni pratiche):

$$\beta_k = \text{rank}(H_k(K; Z))$$

Il significato geometrico dei numeri di Betti è il seguente. β_0 è il numero di componenti connesse del complesso: conta le parti disconnesse dello spazio. β_1 è il numero di cicli indipendenti (buchi unidimensionali, o «loop»): conta le strutture circolari che non sono il bordo di nessuna superficie nel complesso. β_2 è il numero di cavità bidimensionali indipendenti: conta le sfere vuote (superfici chiuse senza riempimento tridimensionale) presenti nel complesso. In generale, β_k conta i «buchi k -dimensionali» del complesso.

Per il reticolo degli atomi della firma informazionale, i numeri di Betti hanno le seguenti interpretazioni semantiche:

$\beta_0(\mathcal{R}, \epsilon) =$ numero di cluster di atomi semanticamente coesi alla scala ϵ . Al variare di ϵ , β_0 parte da $|\mathcal{A}|$ (tutti gli atomi isolati) e converge a 1 (tutto il reticolo connesso) al crescere di ϵ . La sequenza di valori ϵ a cui β_0 decresce di 1 (per unione di due componenti connesse) traccia la struttura gerarchica della coesione semantica nel reticolo.

$\beta_1(\mathcal{R}, \epsilon) =$ numero di cicli informativi indipendenti alla scala ϵ . Cicli con $\beta_1 > 0$ indicano strutture circolari di dipendenza informativa: regioni del reticolo in cui ci sono percorsi multipli tra atomi e in cui non tutti i percorsi sono equivalenti. Nella teoria del contenuto, questi cicli corrispondono a forme di coerenza strutturale che non si lasciano ridurre a semplici catene di dipendenze.

$\beta_2(\mathcal{R}, \epsilon) =$ numero di cavità bidimensionali indipendenti. Cavità con $\beta_2 > 0$ indicano regioni dello spazio informazionale che sono «circondate» da atomi ma non «riempite»: strutture di dipendenza che delimitano una zona dello spazio informazionale senza che quella zona sia direttamente rappresentata. Nella teoria del contenuto, queste cavità corrispondono a zone di incertezza strutturata, aspetti del dominio che il sistema

riconosce come importanti (perché delimitati da atomi presenti) senza rappresentarli direttamente.

V.6,4 — Il diagramma di persistenza e il codice a barre

L'omologia persistente non studia il complesso $VR(\mathcal{A}, \epsilon)$ a un singolo valore di ϵ , ma traccia come i gruppi di omologia evolvono al variare di ϵ attraverso la filtrazione. Il risultato è catturato dal diagramma di persistenza $DP_k(\mathcal{R})$ per ciascun grado k , che è un multiinsieme di coppie $(b_i, d_i) \in \mathbb{R}^2$ con $b_i < d_i$. Ogni coppia corrisponde a un generatore omologico (una componente connessa, un ciclo, una cavità) che nasce a scala b_i e muore a scala d_i . La persistenza del generatore è $\text{pers}_i = d_i - b_i$.

La struttura del diagramma di persistenza risponde al Teorema di Stabilità dell'Omologia Persistente (Edelsbrunner-Harer 2008, Chazal et al. 2012): se due spazi metrici X e Y sono vicini nel senso della distanza di Gromov-Hausdorff, allora i loro diagrammi di persistenza sono vicini nel senso della distanza di bottleneck. La distanza di bottleneck tra diagrammi di persistenza D e D' è:

$$d_B(D, D') = \inf_{\gamma} \sup_{\{x \in D\}} \|x - \gamma(x)\|_{\infty}$$

dove l'infimo è preso su tutte le biiezioni γ tra D e D' aumentati con i punti della diagonale (ciascun punto fuori dalla diagonale può essere «abbinato» a qualsiasi punto della diagonale $b = d$). Questa proprietà di stabilità è cruciale per l'applicabilità pratica: garantisce che piccole perturbazioni nella struttura metrica del reticolo degli atomi producano piccole perturbazioni nei diagrammi di persistenza, il che rende i diagrammi robusti rispetto a errori di misura e variazioni individuali nelle firme.

Una rappresentazione alternativa e spesso più leggibile del diagramma di persistenza è il codice a barre (barcode) $BC_k(\mathcal{R})$, che è l'insieme degli intervalli $[b_i, d_i)$ per ciascun generatore, visualizzabili come barre orizzontali su un asse della scala ϵ . Le barre lunghe corrispondono a strutture topologiche robuste e ad alta persistenza; le barre corte corrispondono a strutture effimere che appaiono e scompaiono a scale simili e sono spesso interpretabili come rumore.

V.6,5 — Calcolo esplicito su un esempio costruito: il reticolo tri-sorgente

Per rendere concrete queste definizioni, si costruisce un esempio esplicito su un sistema con tre sorgenti X_1, X_2, X_3 e un obiettivo Y . In questo caso, il reticolo degli atomi PID ha una struttura che è stata caratterizzata con precisione da Williams e Beer e poi da Bertschinger e colleghi. Per $n=3$ sorgenti, il numero di atomi nella decomposizione PID è 18 (nel caso della misura $I_{\min}$ di Williams-Beer) o varia con la scelta

dell'assioma di non-negatività. Per semplicità didattica, si considererà la decomposizione ridotta ai 7 atomi principali dell'approccio di Williams-Beer nel caso in cui i tre contributi di ordine superiore sono non-nulli:

$\alpha_1 = \{1\}\{2\}\{3\}$ (ridondanza di terzo ordine: informazione condivisa da tutte e tre le sorgenti) $\alpha_2 = \{1\}\{2\}$ (ridondanza tra X_1 e X_2 , non condivisa con X_3) $\alpha_3 = \{1\}\{3\}$ (ridondanza tra X_1 e X_3 , non condivisa con X_2) $\alpha_4 = \{2\}\{3\}$ (ridondanza tra X_2 e X_3 , non condivisa con X_1) $\alpha_5 = \{1\}$ (informazione unica di X_1) $\alpha_6 = \{2\}$ (informazione unica di X_2) $\alpha_7 = \{3\}$ (informazione unica di X_3)

A questi 7 atomi si aggiungono le componenti sinergiche: $\alpha_8 = \{12\}$ (sinergia tra X_1 e X_2) $\alpha_9 = \{13\}$ (sinergia tra X_1 e X_3) $\alpha_{10} = \{23\}$ (sinergia tra X_2 e X_3) $\alpha_{11} = \{123\}$ (sinergia di terzo ordine)

Consideriamo un sistema concreto con la seguente struttura di dipendenze informatiche, scelta per illustrare una configurazione interessante non-generica. Sia Y una variabile binaria ($Y \in \{0, 1\}$) e siano X_1, X_2, X_3 tre sorgenti con la seguente distribuzione congiunta $P(X_1, X_2, X_3, Y)$:

P costruita secondo lo schema seguente. X_1 è una copia rumorosa di Y con probabilità di errore $p_1 = 0,1$ (canale binario simmetrico con flip probability 0,1). X_2 è una copia rumorosa di Y con probabilità di errore $p_2 = 0,15$. X_3 ha un meccanismo diverso: $X_3 = X_1 \text{ XOR } X_2$ (con probabilità di errore $p_3 = 0,05$ dovuta a rumore indipendente). In questo sistema, X_1 e X_2 portano ridondanza rispetto a Y (entrambe indicano direttamente Y), mentre X_3 porta informazione sinergica rispetto alla coppia $\{X_1, X_2\}$ (perché $X_3 = X_1 \text{ XOR } X_2$ fornisce informazione sull'errore congiunto della coppia).

Calcoliamo le distribuzioni condizionali $P_{\alpha_i}(Y)$ per ciascun atomo. Per i valori di entropia e informazione mutua con $p_1 = 0,1$, $p_2 = 0,15$, $p_3 = 0,05$ e distribuzione uniforme $P(Y=0) = P(Y=1) = 0,5$:

$$H(Y) = 1 \text{ bit}$$

$$I(X_1; Y) = H(Y) - H(Y|X_1) = 1 - h_b(0,1) \approx 1 - 0,469 = 0,531 \text{ bit}$$

dove $h_b(p) = -p \log_2 p - (1-p) \log_2(1-p)$ è l'entropia binaria.

$$I(X_2; Y) = 1 - h_b(0,15) \approx 1 - 0,610 = 0,390 \text{ bit}$$

Per $X_3 = X_1 \text{ XOR } X_2$ (più rumore indipendente), il canale da Y a X_3 ha probabilità di errore effettiva $p_{\text{eff}} = p_1(1-p_2) + p_2(1-p_1) + p_3(1 - p_1 - p_2 + 2p_1p_2) \approx 0,1 \cdot 0,85 + 0,15 \cdot 0,9 + \text{small correction} \approx 0,085 + 0,135 \approx 0,220$.

$$I(X_3; Y) = 1 - h_b(0,220) \approx 1 - 0,773 = 0,227 \text{ bit}$$

$$I(X_1, X_2; Y) = H(Y) - H(Y | X_1, X_2)$$

Per calcolare $H(Y | X_1, X_2)$: dato $X_1 = x_1$ e $X_2 = x_2$, la probabilità a posteriori di Y dipende da entrambi gli input. Se $x_1 = x_2$, la stima di Y è rafforzata; se $x_1 \neq x_2$, le due sorgenti sono in conflitto. Con i parametri scelti:

$$\begin{aligned} P(Y=0 | X_1=0, X_2=0) &= P(X_1=0, X_2=0 | Y=0) \cdot P(Y=0) / P(X_1=0, X_2=0) \\ &= (0,9)(0,85)(0,5) / [(0,9)(0,85)(0,5) + (0,1)(0,15)(0,5)] = 0,3825 / (0,3825 + 0,0075) = 0,3825 / 0,3900 \approx 0,981 \end{aligned}$$

$$\begin{aligned} P(Y=0 | X_1=1, X_2=1) &= P(X_1=1, X_2=1 | Y=0) \cdot P(Y=0) / P(X_1=1, X_2=1) \\ &= (0,1)(0,15)(0,5) / [(0,1)(0,15)(0,5) + (0,9)(0,85)(0,5)] = 0,0075 / 0,3900 \approx 0,019 \end{aligned}$$

$$\begin{aligned} P(Y=0 | X_1=0, X_2=1) &= P(X_1=0, X_2=1 | Y=0) \cdot P(Y=0) / P(X_1=0, X_2=1) \\ &= (0,9)(0,15)(0,5) / [(0,9)(0,15)(0,5) + (0,1)(0,85)(0,5)] = 0,0675 / (0,0675 + 0,0425) = 0,0675 / 0,11 \approx 0,614 \end{aligned}$$

I casi $X_1=0, X_2=0$ e $X_1=1, X_2=1$ contribuiscono $H(Y | X_1=0, X_2=0) = h_b(0,019) \approx 0,126$ bit e $H(Y | X_1=1, X_2=1) \approx 0,126$ bit. I casi conflitto $X_1=0, X_2=1$ e $X_1=1, X_2=0$ contribuiscono $H(Y | \text{conflitto}) = h_b(0,614) \approx 0,962$ bit. Con pesi proporzionali alla probabilità marginale di ciascuna configurazione:

$$\begin{aligned} P(X_1=0, X_2=0) &= P(X_1=0, X_2=0 | Y=0) \cdot P(Y=0) + \\ &P(X_1=0, X_2=0 | Y=1) \cdot P(Y=1) = 0,9 \cdot 0,85 \cdot 0,5 + 0,1 \cdot 0,15 \cdot 0,5 = 0,3825 + 0,0075 = 0,39 \end{aligned}$$

$$\text{Analogamente } P(X_1=1, X_2=1) = 0,39, P(X_1=0, X_2=1) = P(X_1=1, X_2=0) = 0,11.$$

$$\begin{aligned} H(Y | X_1, X_2) &= 0,39 \cdot 0,126 + 0,39 \cdot 0,126 + 0,11 \cdot 0,962 + 0,11 \cdot 0,962 = \\ &2 \cdot (0,39)(0,126) + 2 \cdot (0,11)(0,962) = 0,09828 + 0,21164 = 0,3099 \text{ bit} \end{aligned}$$

$$I(X_1, X_2; Y) = 1 - 0,3099 = 0,690 \text{ bit}$$

Per la decomposizione PID, la ridondanza $R(X_1, X_2; Y)$ nel senso di Williams-Beer è definita come:

$$R(X_1, X_2; Y) = I_{\min}(X_1, X_2; Y) = \min(I(X_1; Y), I(X_2; Y)) = \min(0,531, 0,390) = 0,390 \text{ bit}$$

$$\begin{aligned} \text{L'informazione unica di } X_1 \text{ rispetto a } X_2 \text{ per } Y \text{ è: } U(X_1; Y \setminus X_2) &= I(X_1; Y) - R(X_1, X_2; Y) = 0,531 - 0,390 = 0,141 \text{ bit} \end{aligned}$$

$$\begin{aligned} \text{L'informazione unica di } X_2: U(X_2; Y \setminus X_1) &= I(X_2; Y) - R(X_1, X_2; Y) = 0,390 - 0,390 = 0,000 \text{ bit} \end{aligned}$$

$$\begin{aligned} \text{La sinergia di ordine 2 tra } X_1 \text{ e } X_2: S(X_1, X_2; Y) &= I(X_1, X_2; Y) - I(X_1; Y) - I(X_2; Y) + R(X_1, X_2; Y) = 0,690 - 0,531 - 0,390 + 0,390 = 0,159 \text{ bit} \end{aligned}$$

Verifica: $R + U(X_1) + U(X_2) + S = 0,390 + 0,141 + 0,000 + 0,159 = 0,690$ ✓ (uguale a $I(X_1, X_2; Y)$)

V.6,6 — Calcolo delle distanze tra atomi e costruzione della filtrazione

Con i pesi degli atomi calcolati nella sezione precedente, si può ora calcolare la metrica tra atomi. Le distribuzioni condizionali associate a ciascun atomo vengono stimate dalla struttura del sistema PID. Per semplicità, utilizziamo una metrica proxy basata sulle covarianze degli atomi rispetto all'obiettivo Y , che è calcolabile più direttamente dai dati informativi.

Sia $\omega(\alpha)$ il peso dell'atomo α nella decomposizione PID. La distribuzione condizionale dell'atomo è stimata attraverso la risposta dell'atomo al condizionamento sull'obiettivo. Per i 7 atomi principali del nostro esempio, le distanze rilevanti sono:

Gli atomi di ridondanza $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ hanno tutti distribuzioni condizionali che puntano verso le stime di massima verosimiglianza di Y , e la distanza tra due atomi di ridondanza è proporzionale alla differenza nei loro pesi: $d(\alpha_i, \alpha_j) \propto |\omega(\alpha_i) - \omega(\alpha_j)| / \max(\omega)$.

L'atomo di sinergia $\alpha_8 = \{12\}$ ha una distribuzione condizionale qualitativamente diversa dagli atomi di ridondanza: risponde positivamente al caso $X_1=X_2=Y$ e negativamente al caso di conflitto $X_1 \neq X_2$. La distanza tra α_8 e gli atomi di ridondanza è quindi più grande della distanza tra atomi di ridondanza tra loro: $d(\alpha_8, \alpha_1) > d(\alpha_1, \alpha_2)$.

Per costruire la filtrazione esplicita, assegnamo le seguenti distanze tipiche (in bit, come misura naturale per le distanze nel nostro spazio informativo):

$d(\alpha_1, \alpha_2) = 0,12$ (ridondanza di ordine 3 vs. ridondanza tra X_1X_2 : simili, entrambe riflettono correlazione con Y) $d(\alpha_1, \alpha_3) = 0,18$ (ridondanza di ordine 3 vs. ridondanza tra X_1X_3 : distanza moderata) $d(\alpha_1, \alpha_4) = 0,22$ (ridondanza di ordine 3 vs. ridondanza tra X_2X_3) $d(\alpha_2, \alpha_3) = 0,20$ (ridondanze parziali con X_1 in comune) $d(\alpha_2, \alpha_4) = 0,28$ (ridondanze parziali senza sorgenti in comune) $d(\alpha_3, \alpha_4) = 0,26$ (ridondanze parziali con X_3 in comune) $d(\alpha_5, \alpha_1) = 0,35$ (informazione unica X_1 vs. ridondanza totale) $d(\alpha_6, \alpha_1) = 0,42$ (informazione unica X_2 vs. ridondanza totale) $d(\alpha_7, \alpha_1) = 0,48$ (informazione unica X_3 vs. ridondanza totale) $d(\alpha_5, \alpha_6) = 0,38$ (informazioni uniche di sorgenti diverse) $d(\alpha_5, \alpha_7) = 0,44$ $d(\alpha_6, \alpha_7) = 0,40$ $d(\alpha_8, \alpha_1) = 0,55$ (sinergia X_1X_2 vs. ridondanza totale: distanza elevata, distribuzioni diverse) $d(\alpha_8, \alpha_2) = 0,50$ (sinergia X_1X_2 vs. ridondanza X_1X_2 : moderate, stesso sottosistema) $d(\alpha_8, \alpha_5) = 0,45$ (sinergia X_1X_2 vs. unica X_1) $d(\alpha_8, \alpha_6) = 0,45$ (sinergia X_1X_2 vs. unica X_2) $d(\alpha_{11}, \alpha_8) = 0,30$ (sinergia totale vs. sinergia di ordine 2: simili per tipo)

$d(\alpha_{11}, \alpha_9) = 0,32$ (sinergia totale vs. sinergia X_1X_3) $d(\alpha_{11}, \alpha_{10}) = 0,35$ (sinergia totale vs. sinergia X_2X_3)

La filtrazione VR a scale $\epsilon = \{0,12, 0,18, 0,22, 0,26, 0,28, 0,30, 0,35, 0,40, 0,45, 0,50, 0,55\}$ produce la seguente sequenza di complessi.

A $\epsilon = 0$: VR₀ ha 11 vertici isolati, nessun spigolo. $\beta_0 = 11, \beta_1 = \beta_2 = 0$.

A $\epsilon = 0,12$: lo spigolo (α_1, α_2) viene incluso. Le componenti di α_1 e α_2 si fondono. $\beta_0 = 10, \beta_1 = \beta_2 = 0$.

A $\epsilon = 0,18$: lo spigolo (α_1, α_3) viene incluso. Siamo ora nella stessa componente $\{\alpha_1, \alpha_2, \alpha_3\}$. $\beta_0 = 9$.

A $\epsilon = 0,20$: lo spigolo (α_2, α_3) viene incluso. Il triangolo $(\alpha_1, \alpha_2, \alpha_3)$ potrebbe formarsi se tutti e tre gli spigoli sono presenti: sì, lo sono. Viene incluso il triangolo. $\beta_0 = 9, \beta_1 = 0$ (il ciclo che sembrava formarsi con $(\alpha_1, \alpha_2), (\alpha_2, \alpha_3), (\alpha_1, \alpha_3)$ è immediatamente riempito dal triangolo). No cambiamento in β_1 .

A $\epsilon = 0,22$: lo spigolo (α_1, α_4) viene incluso. α_4 si aggrega al cluster $\{\alpha_1, \alpha_2, \alpha_3\}$. $\beta_0 = 8$.

A $\epsilon = 0,26$: lo spigolo (α_3, α_4) viene incluso. Ora abbiamo il ciclo $(\alpha_1, \alpha_3, \alpha_4)$ ma manca ancora (α_2, α_4) . Il quadrato $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ non è completo: (α_2, α_4) ha distanza $0,28 > 0,26$. Si forma il ciclo $(\alpha_1, \alpha_3, \alpha_4, \alpha_1)$ senza il riempimento. β_1 aumenta a 1.

A $\epsilon = 0,28$: lo spigolo (α_2, α_4) viene incluso. Ora il quadrilatero $(\alpha_1, \alpha_2, \alpha_4, \alpha_3)$ ha tutti i lati. Il ciclo precedente viene chiuso? Il ciclo $\alpha_1 \rightarrow \alpha_3 \rightarrow \alpha_4 \rightarrow \alpha_1$ esiste, e ora con lo spigolo $\alpha_2 \rightarrow \alpha_4$ si forma un nuovo path. Il ciclo $\alpha_1 \rightarrow \alpha_2 \rightarrow \alpha_4 \rightarrow \alpha_3 \rightarrow \alpha_1$ non è ancora chiuso perché manca (α_2, α_3) ... no, aspetto: a $\epsilon=0,20$ avevamo già (α_2, α_3) . Quindi a $\epsilon=0,28$ il quadrilatero $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ ha tutti e sei gli spigoli delle sue 4 coppie: $(\alpha_1, \alpha_2), (\alpha_1, \alpha_3), (\alpha_1, \alpha_4), (\alpha_2, \alpha_3), (\alpha_2, \alpha_4), (\alpha_3, \alpha_4)$. Il complesso VR include il tetraedro pieno $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$. Il ciclo $\beta_1=1$ che si era formato a $\epsilon=0,26$ viene ora riempito. β_1 scende a 0. β_0 rimane 8 (poiché α_2 e α_4 erano già nella stessa componente di α_1).

A $\epsilon = 0,30$: lo spigolo (α_{11}, α_8) viene incluso. α_{11} e α_8 si connettono. $\beta_0 = 7$ (se α_{11} e α_8 erano in componenti distinte, cosa che è perché non erano state collegate prima).

A $\epsilon = 0,32$: lo spigolo (α_{11}, α_9) viene incluso. α_9 si aggiunge al cluster $\{\alpha_{11}, \alpha_8\}$. $\beta_0 = 6$.

A $\epsilon = 0,35$: vengono inclusi simultaneamente $(\alpha_{11}, \alpha_{10}), (\alpha_5, \alpha_1)$, e altri spigoli. Il cluster delle sinergie $\{\alpha_8, \alpha_9, \alpha_{10}, \alpha_{11}\}$ si forma completamente e inizia a connettersi con il cluster delle ridondanze $\{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$. Il

cluster delle informazioni uniche $\{\alpha_5\}$ si connette a α_1 . $\beta_0 = 3$ (approssimativamente, dopo queste fusioni).

A $\varepsilon = 0,40$: α_6 e α_7 si connettono agli altri cluster. $\beta_0 = 1$ (tutto il reticolo è connesso). A questo punto β_1 potrebbe essere non nulla se si sono formati cicli senza riempimento nel processo di fusione.

A $\varepsilon = 0,45$: gli spigoli tra il cluster delle sinergie e il cluster delle informazioni uniche vengono inclusi. Possibili formazioni di nuovi cicli o loro chiusura.

A $\varepsilon = 0,55$: il reticolo completo è quasi pieno. Tutti gli spigoli rimanenti vengono inclusi.

V.6,7 — Il codice a barre di persistenza per il reticolo tri-sorgente

Dalla filtrazione costruita nella sezione precedente, il codice a barre di persistenza $BC_k(\mathcal{R})$ in ciascun grado k è il seguente.

Il codice a barre di grado 0 (componenti connesse):

La storia delle fusioni può essere rappresentata come un dendrogramma. Partendo da 11 componenti a $\varepsilon=0$, le barre di persistenza in grado 0 hanno le seguenti nascite e morti:

$[0, 0,12)$: α_1 come componente isolata. Nasce a 0, muore a 0,12 (quando si fonde con α_2). $[0, 0,12)$: α_2 come componente isolata. Nasce a 0, muore a 0,12. $[0, 0,18)$: α_3 come componente isolata. Nasce a 0, muore a 0,18. $[0, 0,22)$: α_4 come componente isolata. Nasce a 0, muore a 0,22. $[0, 0,35)$: α_5 come componente isolata. Nasce a 0, muore a 0,35. $[0, 0,35)$: α_8 come componente isolata. Nasce a 0, muore a 0,30. $[0, 0,40)$: α_6 come componente isolata. $[0, 0,40)$: α_7 come componente isolata. $[0, \infty)$: il cluster principale. Una componente nasce a 0 e persiste per tutta la filtrazione (la componente «sopravvissuta» è quella che non muore mai).

In grado 0 si ha sempre esattamente una barra $[0, \infty)$ che rappresenta la componente sopravvissuta (il componente connessa principale). Tutte le altre barre hanno durata finita.

Il codice a barre di grado 1 (cicli):

La barra più significativa è quella corrispondente al ciclo che nasce a $\varepsilon=0,26$ (quando si forma il ciclo tra gli atomi di ridondanza $\alpha_1, \alpha_3, \alpha_4$ senza ancora lo spigolo $\alpha_2-\alpha_4$) e muore a $\varepsilon=0,28$ (quando viene riempito). Questa barra ha persistenza $\text{pers} = 0,28 - 0,26 = 0,02$ bit, che è molto bassa: il ciclo è effimero e probabilmente è artefatto della struttura discreta delle distanze.

Potrebbero esistere cicli a scala maggiore tra i cluster delle ridondanze e delle sinergie, che nascono quando i due cluster iniziano a connettersi a scale diverse e producono percorsi multipli tra la regione delle ridondanze e la regione delle sinergie. Se esiste uno spigolo breve tra α_8 (distanza 0,50) e uno spigolo lungo tra α_1 e α_{11} (distanza ipotetica 0,52), potrebbe formarsi un ciclo che nasce a 0,50 e muore a 0,52, con persistenza 0,02, anch'essa molto bassa.

L'assenza di cicli ad alta persistenza in questo esempio riflette una struttura informazionale «tree-like» del sistema: i due cluster principali (ridondanze e sinergie) si connettono attraverso un'unica bottleneck informazionale, senza percorsi multipli indipendenti. Questa struttura ha un'interpretazione semantica precisa: l'informazione ridondante e l'informazione sinergica del sistema si raccordano attraverso un'unica interfaccia funzionale, il che rende il sistema vulnerabile al taglio di quella interfaccia ma anche molto efficiente nella trasmissione dell'informazione dal livello ridondante al livello sinergico.

Il codice a barre di grado 2 (cavità):

Per questo sistema tri-sorgente, $\beta_2 = 0$ per tutte le scale ϵ , il che indica l'assenza di cavità tridimensionali. Questo risultato è coerente con la struttura relativamente semplice del sistema: con solo 11 atomi e un numero moderato di dipendenze, il reticolo non genera strutture tridimensionali vuote. Sistemi più complessi — con $n > 4$ sorgenti e strutture di dipendenza più ricche — possono avere $\beta_2 > 0$, e le cavità risultanti hanno interpretazioni semantiche importanti come zone di incertezza strutturata.

V.6,8 — Invarianti di Betti della firma informazionale: definizione e proprietà

L'analisi della sezione precedente ha mostrato come i numeri di Betti varino con la scala ϵ nella filtrazione. La teoria dell'omologia persistente fornisce un modo per estrarre invarianti topologici globali che riassumono l'intera struttura del codice a barre in singoli numeri caratteristici. Definiamo tre invarianti di Betti della firma informazionale.

Il primo invariante è il numero di Betti β_0^* (pronunciato «beta-zero-stella»), definito come il numero di barre nel codice a barre di grado 0 con persistenza superiore a una soglia θ :

$$\beta_0^*(\Sigma, \theta) = |\{ (b_i, d_i) \in BC_0(\mathcal{R}) : d_i - b_i > \theta \}|$$

Questo numero conta le componenti connesse «robuste» della firma — cluster di atomi che rimangono distinti per un intervallo di scala di ampiezza almeno θ prima di fondersi con altri cluster. Il parametro θ funge da filtro del rumore: cluster effimeri (che si formano e fondono rapidamente al variare di ϵ) non vengono conteggiati. Per $\theta = 0,10$ bit nel nostro esempio, $\beta_0^* = 3$: i cluster delle ridondanze di terzo ordine, delle

ridondanze di secondo ordine, e il cluster delle sinergie sono robusti e si fondono solo a scale grandi.

Il secondo invariante è il numero di Betti β_1^* (cicli robusti):

$$\beta_1^*(\Sigma, \theta) = |\{ (b_i, d_i) \in BC_1(\mathcal{R}) : d_i - b_i > \theta \}|$$

Conta i cicli informativi che persistono per almeno θ unità di scala prima di essere riempiti. Nel nostro esempio, $\beta_1^* = 0$ per $\theta = 0,05$ (nessun ciclo ha persistenza superiore a 0,02). Un sistema con $\beta_1^* > 0$ ha strutture circolari robuste di dipendenza informativa: regioni del reticolo in cui ci sono percorsi multipli non equivalenti tra atomi, indicativi di ridondanza strutturale di tipo «circolare» che va oltre la semplice ridondanza lineare catturata dagli atomi PID di basso ordine.

Il terzo invariante è il numero di Betti ponderato dalla persistenza β_1 :

$$\beta_1(\Sigma) = \sum_i (d_i - b_i) \text{ per } (b_i, d_i) \in BC_1(\mathcal{R})$$

Questo invariante somma le persistenze di tutti i cicli, dando un peso maggiore ai cicli più robusti. È una misura scalare della complessità ciclica del reticolo degli atomi. Nel nostro esempio, $\beta_1 \approx 0,02$ (contributo del solo ciclo effimero). In sistemi più complessi, β_1 può essere sostanziale, indicando una struttura ciclica ricca.

Un quarto invariante, particolarmente utile per confronti tra firme di sistemi diversi, è la norma del diagramma di persistenza in grado 1, definita come la distanza alla diagonale del diagramma:

$$\|DP_1(\Sigma)\|_p = (\sum_i (\text{pers}_i)^p)^{1/p}$$

Per $p=2$, questa è la norma euclidea delle persistenze dei cicli. Questa norma è stabile rispetto a perturbazioni della metrica del reticolo (per il Teorema di Stabilità citato sopra), il che la rende adatta come misura di complessità topologica robusta.

Le proprietà formali degli invarianti di Betti della firma informazionale sono le seguenti.

Stabilità: Per il Teorema di Stabilità dell'Omologia Persistente, se due firme Σ e Σ' hanno distanza $d_{GH}(\Sigma, \Sigma') \leq \delta$, allora i loro diagrammi di persistenza hanno distanza di bottleneck $d_B(DP_k(\Sigma), DP_k(\Sigma')) \leq \delta$ per ogni k . Questa proprietà garantisce che gli invarianti di Betti varino continuamente con la firma.

Invarianza rispetto all'isometria: Se Σ e Σ' sono due firme i cui reticoli degli atomi sono isometrici (c'è una biiezione che preserva tutte le distanze tra atomi), allora i loro diagrammi di persistenza sono identici. Gli invarianti di Betti catturano dunque proprietà intrinseche della struttura metrica del reticolo, non dipendenti dalla numerazione o dall'ordine degli atomi.

Completezza parziale: I diagrammi di persistenza non sono completi nel senso che due firme con gli stessi diagrammi di persistenza non sono necessariamente isometriche. Esistono informazioni sulla struttura metrica del reticolo che non sono catturate dai diagrammi di persistenza. Tuttavia, per molti scopi pratici, i diagrammi forniscono una caratterizzazione sufficientemente ricca da distinguere firme strutturalmente diverse.

V.6,9 — Interpretazione semantica degli invarianti topologici

Con le definizioni e i calcoli delle sezioni precedenti, si può ora fornire un'interpretazione semantica sistematica degli invarianti topologici della firma informazionale.

Il numero $\beta_0^*(\Sigma, \theta)$ è una misura della modularità semantica del contenuto: conta le componenti semantiche robuste che compongono il contenuto. Un contenuto con alta modularità semantica (β_0^* grande) è un contenuto in cui le diverse componenti informative sono relativamente indipendenti e si possono «dissociare» senza perdere le altre. Questo corrisponde alla struttura di contenuti come «il volto di una persona sconosciuta» (dove le componenti fisiognomica, emotiva ed episodica sono relativamente indipendenti) al contrario di contenuti come «il volto di mia madre» (dove le componenti sono altamente correlate, come mostrano le dissociazioni patologiche nella prosopagnosia).

Il numero $\beta_1^*(\Sigma, \theta)$ è una misura della struttura ciclica del contenuto: conta le dipendenze informative circolari robuste. Un contenuto con alta struttura ciclica (β_1^* grande) è un contenuto in cui la struttura informazionale ha retroazioni multiple: l'informazione su un aspetto del contenuto si propaga attraverso percorsi multipli non equivalenti, raggiungendo lo stesso punto di partenza da direzioni diverse. Questo corrisponde a strutture semantiche come il frame della causalità (dove causa ed effetto si definiscono circolarmente) o le strutture di auto-riferimento nel linguaggio.

La norma $\|DP_1(\Sigma)\|$ è una misura della complessità topologica globale del contenuto. Contenuti con alta complessità topologica sono contenuti con strutture di dipendenza ricche e intricate, che richiedono l'elaborazione di percorsi informativi multipli per essere compresi appieno. Questa misura è correlata con la difficoltà cognitiva di un contenuto: contenuti topologicamente complessi richiedono più risorse cognitive per essere elaborati, memorizzati e comunicati.

V.6,10 — La topologia persistente come strumento per il confronto tra specie e tra individui

L'applicazione più promettente degli invarianti topologici della firma informazionale è quella del confronto sistematico tra sistemi cognitivi diversi: tra individui diversi, tra specie diverse, tra stati di sviluppo diversi dello stesso individuo. Il confronto è reso possibile dalla proprietà di stabilità degli invarianti: piccole differenze nelle firme producono piccole differenze negli invarianti, il che garantisce che il confronto sia robusto rispetto alla variabilità nei dati.

Un confronto sistematico tra gli invarianti di Betti delle firme di molti soggetti rispetto allo stesso contenuto fornisce una caratterizzazione della variabilità interindividuale della struttura topologica del contenuto. Se $\beta_0^*(\Sigma_A, \theta) \neq \beta_0^*(\Sigma_B, \theta)$ per soggetti A e B rispetto allo stesso contenuto, i due soggetti hanno strutture modulari diverse del contenuto: hanno componenti robuste diverse, il che indica che il contenuto si organizza in modo qualitativamente diverso nei due sistemi. Se $\beta_1^*(\Sigma_A, \theta) > \beta_1^*(\Sigma_B, \theta)$, il soggetto A ha strutture di dipendenza circolare più ricche, il che suggerisce una rappresentazione più integrata e interconnessa del contenuto.

Per il confronto tra specie, l'analisi topologica può rivelare differenze strutturali nel modo in cui diverse specie rappresentano concetti condivisi. Se i roditori e i primati hanno firme informazionali per il concetto «predatore» con diversi β_0^* (diversa modularità semantica), questo indica che il contenuto «predatore» ha una struttura qualitativamente diversa nelle due specie: diverse componenti informazionali robuste, diversa organizzazione modulare. Queste differenze strutturali hanno implicazioni per i comportamenti difensivi: una struttura con β_0^* maggiore (più modularità) permette dissociazioni selettive (ad esempio, risposta alla forma ma non all'odore, o vice versa), mentre una struttura con β_0^* minore (più integrazione) produce risposte più olistiche che attivano l'intero pattern difensivo anche in presenza di segnali parziali.

V.6,11 — Relazione tra la topologia persistente e le altre misure della firma

Per completare la trattazione, è opportuno collocare gli invarianti topologici nel sistema delle misure della firma informazionale già introdotte nei capitoli precedenti.

Il numero di atomi robusti $\kappa(\Sigma, \theta)$ — introdotto nel Capitolo IV come misura della ricchezza rappresentazionale — è correlato ma non identico a $\beta_0^*(\Sigma, \theta)$. κ conta gli atomi con peso superiore a una soglia, mentre β_0^* conta i cluster di atomi che rimangono distinti per un range di scale. Un sistema con κ grande ma $\beta_0^* = 1$ ha molti atomi ricchi ma tutti organizzati in un unico cluster senza struttura modulare: è un sistema con contenuto ricco ma semanticamente indifferenziato. Un sistema con κ moderato e β_0^* grande ha pochi atomi ma ben separati in componenti distinte: è un sistema con contenuto semanticamente strutturato anche se

non particolarmente ricco. La misura di ricchezza rappresentazionale $\varrho(\Xi) = \kappa(\Sigma) \cdot (1 + \sigma(\Sigma))$ cattura sia la ricchezza degli atomi sia la presenza di sinergie, ma non la struttura topologica dei cluster: è possibile avere ϱ alta e $\beta_1^* = 0$ (ricchezza senza struttura ciclica), o ϱ moderata e $\beta_1^* > 0$ (struttura ciclica senza alta ricchezza).

La distanza di Gromov-Hausdorff $d_{GH}(\Sigma, \Sigma')$ cattura la distanza globale tra firme, ma è sensibile alla struttura locale delle relazioni tra atomi. La distanza di bottleneck $d_B(DP_k(\Sigma), DP_k(\Sigma'))$ tra i diagrammi di persistenza è meno sensibile alla struttura locale e più sensibile alla struttura topologica globale. Per molti scopi teorici, la distanza di bottleneck è una misura più robusta della distanza tra contenuti, perché filtra le differenze locali irrilevanti e cattura le differenze nella struttura topologica di ordine superiore.

Il grado di oggettività $\Omega(\Theta)$ — definito come la dispersione delle firme tra i sistemi della famiglia di riferimento — può essere raffinato usando la distanza di bottleneck al posto della distanza di Gromov-Hausdorff: la versione topologica del grado di oggettività misura quanto le strutture topologiche delle firme siano allineate tra i sistemi, indipendentemente da differenze locali nella metrica degli atomi.

V.6,12 — Verso la topologia persistente multivalore: il caso dei contenuti dinamici

Un'estensione importante degli strumenti discussi nelle sezioni precedenti è quella della topologia persistente multivalore, in cui la filtrazione non è parametrizzata da un singolo parametro ε ma da un vettore di parametri $(\varepsilon_1, \dots, \varepsilon_m)$. Questo caso è rilevante per la firma informazionale perché le distanze tra atomi possono essere calcolate rispetto a diverse dimensioni semantiche: la distanza cognitiva (quanto due atomi co-variano nell'elaborazione cognitiva), la distanza evolutiva (quanto le funzioni proprie degli atomi sono evolutivamente connesse), la distanza temporale (quanto i due atomi co-variano nel tempo). Ciascuna di queste dimensioni definisce una diversa metrica sul reticolo degli atomi, e la struttura topologica può essere diversa nelle diverse dimensioni.

L'omologia persistente multivalore (o omologia persistente a parametri multipli) è uno strumento formalmente più ricco ma anche computazionalmente più impegnativo dell'omologia a parametro singolo. Il suo sviluppo teorico è ancora parziale: il teorema di stabilità dell'omologia a parametro singolo non si generalizza direttamente al caso multivalore, e la teoria dei moduli di persistenza a parametri multipli è attiva area di ricerca. Tuttavia, anche in assenza di una teoria completa, il concetto di filtrazione multivalore offre un modo di pensare alla struttura topologica della firma informazionale che è più fedele alla complessità reale del fenomeno: il contenuto ha struttura topologica diversa a diverse

scale semantiche, e la struttura topologica a una scala non determina univocamente quella alle altre.

Per il presente scopo, è sufficiente notare che la filtrazione multivalore permette di distinguere due tipi di strutture topologiche nella firma informazionale: le strutture che sono robuste a tutte le scale semantiche (strutture che appaiono con $\beta_{k^*} > 0$ in ogni dimensione della filtrazione) e le strutture che sono robuste solo rispetto ad alcune dimensioni. Le prime corrispondono a aspetti del contenuto che sono strutturati in modo semanticamente univoco; le seconde corrispondono ad aspetti del contenuto che sono strutturati in modo diverso a seconda della prospettiva semantica da cui si guardano, e questa molteplicità di strutture è essa stessa semanticamente informativa.

V.7 — Il grado di oggettività del contenuto e la sua misura sperimentale

La combinazione della distanza di Gromov-Hausdorff e del grado di oggettività Ω permette di formulare previsioni empiriche precise sulla struttura della variabilità interindividuale nel contenuto. Se la tesi della ridondanza costitutiva è corretta, si attende che:

Il grado di oggettività $\Omega(\Theta)$ sia correlato positivamente con la robustezza funzionale del contenuto Θ nelle rappresentazioni neurali: i contenuti con firme altamente allineate tra individui diversi dovrebbero essere più stabili rispetto a disturbi locali, più facili da comunicare attraverso il linguaggio, e meno soggetti a variazione nelle valutazioni di somiglianza categoriale.

La distanza $d_{GH}(\Sigma_A(\Theta), \Sigma_B(\Theta))$ tra le firme di due individui A e B rispetto al contenuto Θ dovrebbe essere correlata positivamente con la divergenza nel comportamento di A e B in compiti che richiedono l'elaborazione di Θ : individui con firme molto distanti dovrebbero mostrare comportamenti sistematicamente diversi nelle situazioni in cui Θ è rilevante.

La variazione intraindividuale della firma informazionale nel tempo — la lunghezza della traiettoria $\Sigma(t)$ nel periodo $[t_0, t_1]$ — dovrebbe essere correlata con misure di apprendimento e aggiornamento concettuale: periodi di apprendimento intenso dovrebbero corrispondere a traiettorie più lunghe, periodi di consolidazione a traiettorie più brevi.

Queste previsioni sono, in linea di principio, testabili attraverso una combinazione di paradigmi di neuroimaging (per misurare le firme informazionali delle strutture neurali), paradigmi comportamentali (per misurare le conseguenze funzionali delle differenze tra firme) e paradigmi computazionali (per stimare le strutture di dipendenza informativa nei dati di attivazione neurale). Il Capitolo VI discuterà come

queste previsioni possano essere tradotte in protocolli sperimentali concreti.

Capitolo VI — Errore rappresentazionale, normatività e la firma di riferimento

VI.1 — Il problema dell'errore nella prospettiva della ridondanza costitutiva

Nessuna teoria del contenuto mentale può considerarsi completa senza una risposta soddisfacente al problema dell'errore rappresentazionale. Il problema è, nella sua formulazione più diretta, il seguente: in che senso uno stato mentale può essere sbagliato rispetto al proprio oggetto, cioè in che senso può non corrispondere a ciò che intende rappresentare? La domanda è filosoficamente impegnativa perché la risposta più ovvia — lo stato è sbagliato quando non corrisponde alla realtà — presuppone una nozione di «corrispondenza» che rimane oscura finché non si specifica in cosa consista tale relazione di corrispondenza tra lo stato interno e il fatto esterno. Per una teoria naturalistica del contenuto, che si propone di spiegare il contenuto in termini di relazioni causali, informazionali o funzionali tra stati fisici, la nozione di corrispondenza non è disponibile come primitivo: deve essere costruita a partire da risorse ontologicamente più parche.

Le soluzioni tradizionali al problema dell'errore si dividono, grosso modo, in due famiglie. La prima famiglia — le teorie asimmetriche della dipendenza, di cui la versione di Fodor è l'esemplare più noto — individua la causa «propria» di uno stato tra le sue molte cause possibili attraverso una relazione di dipendenza asimmetrica: la causa propria di uno stato R del tipo T è la causa C_1 tale che ogni altra causa C_2 che produce stati del tipo T è causalmente dipendente da C_1 , ma non viceversa. La misrappresentazione si produce quando uno stato del tipo T è prodotto da una causa diversa da C_1 , che è causalmente dipendente da C_1 ma non ne condivide il contenuto. La seconda famiglia — le teorie funzionali-teleologiche, di cui le versioni di Millikan e Dretske sono i principali esponenti — individua il contenuto attraverso la funzione biologica o informazionale dello stato, e la misrappresentazione si produce quando lo stato si attiva in condizioni in cui la funzione non è correttamente eseguita.

Entrambe queste famiglie incontrano un problema comune nella zona della ridondanza: come identificare la causa «propria» o la funzione

«corretta» quando lo stato in questione ha non una ma molte strutture causali che contribuiscono al suo contenuto? La teoria asimmetrica della dipendenza di Fodor funziona agevolmente per rappresentazioni con un singolo tipo causale di attivazione, ma diventa problematica quando il sistema è costruito per rispondere a un oggetto attraverso molti canali causali diversi e indipendenti — precisamente la struttura caratteristica delle rappresentazioni ridondanti. La teleosemantica di Millikan, come si è discusso nel Capitolo III, incontra la difficoltà di scegliere il livello di descrizione funzionale rilevante in sistemi con architettura reticolare.

La prospettiva della ridondanza costitutiva offre una soluzione che non presuppone l'identificazione di una causa privilegiata o di una funzione privilegiata. L'errore rappresentazionale è caratterizzato come incoerenza della firma: uno stato del sistema Ξ misrappresenta Θ nella misura in cui la sua firma informazionale si discosta dalla firma di riferimento $\Sigma^*(\Theta)$ — la firma che il sistema produce quando l'input è genuinamente un'istanza di Θ in condizioni normali di funzionamento. Non è necessario sapere quale sia la causa propria dello stato; basta confrontare la struttura attuale delle relazioni tra le rappresentazioni ridondanti con la struttura di riferimento.

VI.2 — La firma di riferimento: definizione rigorosa

La nozione di firma di riferimento è il perno dell'intera teoria della normatività basata sulla ridondanza costitutiva, e merita una definizione rigorosa che chiarisca sia cosa essa sia sia cosa non sia.

Sia $\Xi = (\xi_1, \dots, \xi_n)$ il sistema delle rappresentazioni ridondanti di un agente cognitivo A , e sia Θ una proprietà o oggetto del mondo. Denotiamo con $\mathcal{F}(\Theta)$ l'insieme di tutte le firme informazionali $\Sigma(\Xi(t), \Theta)$ che il sistema produce in risposta a input del tipo Θ , al variare del tempo t , del contesto e delle condizioni di stimolazione. La firma di riferimento è:

$$\Sigma^*(\Theta) \equiv \operatorname{argmax}_{\{\Sigma \in \mathcal{F}(\Theta)\}} \kappa(\Sigma)$$

dove κ è il coefficiente di coerenza definito nel Capitolo IV. La firma di riferimento è dunque la firma di massima coerenza nell'insieme delle firme producibili dal sistema in risposta a input del tipo Θ . Essa corrisponde, intuitivamente, alla «migliore rappresentazione possibile» del sistema per quell'oggetto — non una rappresentazione ideale esterna al sistema, ma la struttura che emerge quando tutte le componenti del sistema funzionano in modo integrato e l'input è sufficientemente ricco da attivare tutti i canali.

Questa definizione ha alcune proprietà che vale la pena esplicitare.

La firma di riferimento è interna al sistema nel senso preciso in cui è un attrattore: non è imposta dall'esterno ma emerge dalla struttura delle interazioni tra le componenti quando il sistema è nelle condizioni operative più favorevoli. È analoga all'attrattore di un sistema dinamico — una struttura che il sistema tende verso di sé, senza che sia necessario un agente esterno che la imponga.

La firma di riferimento è normativa nel senso che costituisce lo standard rispetto al quale si valuta ogni stato attuale. Ma la normatività qui non ha carattere trascendente: non è la norma di un sistema ideale che il sistema reale dovrebbe approssimare, ma la norma interna del sistema reale stesso, definita dai suoi stessi stati di massima coerenza. Questo la distingue dalla firma di riferimento nel senso di Millikan — la condizione di correttezza come condizione di proprietà normale — perché la firma di riferimento è definita rispetto alla struttura attuale del sistema, non alla storia selettiva.

La firma di riferimento è dinamica nel senso che cambia quando il sistema apprende. L'insieme $\mathcal{F}(\Theta)$ si espande man mano che il sistema acquisisce nuove rappresentazioni o affina quelle esistenti, e il massimo del coefficiente di coerenza su $\mathcal{F}(\Theta)$ può spostarsi verso valori più elevati. Questo cambiamento è precisamente il progresso cognitivo: il sistema impara a rappresentare Θ con maggiore coerenza.

La misura di errore rappresentazionale $\text{Err}(\Xi, \Theta)$ è definita come la distanza di Gromov-Hausdorff tra la firma attuale e la firma di riferimento:

$$\text{Err}(\Xi, \Theta) = d_{\text{GH}}(\Sigma(\Xi, \Theta), \Sigma^*(\Theta))$$

Questa misura è zero quando il sistema è nella sua configurazione di massima coerenza per il contenuto Θ — quando non vi è errore — e cresce al crescere della discrepanza tra la struttura attuale e la struttura di riferimento. L'errore è dunque una grandezza continua e misurabile, non una proprietà binaria.

VI.3 — Tassonomia degli errori rappresentazionali nella prospettiva della firma

Una delle risorse più utili offerte dal quadro formale della firma informazionale è la possibilità di classificare i tipi di errore rappresentazionale in modo più preciso di quanto le teorie tradizionali consentano. Le teorie standard distinguono tipicamente tra errore per falsa attivazione (il sistema si attiva in assenza del referente), errore per mancata attivazione (il sistema non si attiva in presenza del referente), ed errore per errata specificazione (il sistema si attiva per il referente sbagliato). Questi tipi di errore sono definiti rispetto alla relazione tra un

singolo stato e il suo referente; non catturano la struttura interna dell'errore nell'architettura ridondante.

La tassonomia che emerge dalla prospettiva della firma distingue tra almeno quattro tipi di errore qualitativamente distinti.

Il primo tipo è l'errore per de-coerenza. In questo caso, la firma attuale ha la struttura reticolare corretta — gli atomi giusti sono presenti con i valori approssimativamente giusti — ma il coefficiente di coerenza κ è ridotto rispetto alla firma di riferimento: le componenti del sistema producono informazioni parzialmente incongruenti tra loro, come se ciascuna componente «stesse dicendo qualcosa di leggermente diverso» riguardo al contenuto. Questo tipo di errore si manifesta come incertezza, ambiguità o instabilità della rappresentazione: il sistema non ha una rappresentazione chiara e stabile dell'oggetto, ma una firma «sfocata» in cui le diverse componenti non si rafforzano reciprocamente.

Il secondo tipo è l'errore per soppressione della ridondanza. In questo caso, l'atomo ridondante massimo $\mathcal{R}(\{1\} \wedge \{2\} \wedge \dots \wedge \{n\}; \Theta)$ è significativamente inferiore al valore di riferimento, mentre gli atomi di informazione unica sono relativamente preservati. Questo tipo di errore corrisponde alla situazione in cui il sistema ha molte rappresentazioni che ciascuna porta informazione individuale sul contenuto, ma queste rappresentazioni non si vincolano reciprocamente nel modo caratteristico della firma di riferimento. Il sistema sa «molte cose» sul referente, ma le sue conoscenze sono parzialmente compartimentate — non si integrano nella struttura di coerenza che caratterizza la conoscenza genuina.

Il terzo tipo è l'errore per distorsione sinergica. In questo caso, la componente sinergica della firma — la sinergia di ordine superiore — è presente ma con valore o struttura diversi da quelli di riferimento. Questo tipo di errore si manifesta come distorsione sistematica del contenuto: il sistema non si limita a ignorare alcune informazioni sul referente (come nell'errore per de-coerenza), ma costruisce attivamente un contenuto deformato in cui le relazioni tra le componenti producono inferenze sistematicamente errate. Il caso del delirio di Capgras, discusso nel Capitolo II, è interpretabile come un errore di questo tipo: la componente sinergica della firma del riconoscimento dell'identità personale — quella che emerge dall'integrazione delle informazioni visive con quelle affettive — è alterata in modo tale da produrre l'inferenza «questa persona è un impostore», che non si produrrebbe se la struttura sinergica fosse quella di riferimento.

Il quarto tipo è l'errore per contaminazione. In questo caso, compaiono nella firma attuale atomi informativi che non sono presenti nella firma di riferimento — informazione su oggetti o proprietà diversi da Θ che si introduce nella firma attraverso le connessioni tra componenti del sistema. Questo tipo di errore si manifesta come confusione categoriale o come generalizzazione eccessiva: il sistema tratta l'oggetto come se

avesse proprietà che appartengono a oggetti di tipo diverso, perché le sue componenti sono «contaminate» da strutture di risposta costruite per altri contenuti. In termini formali:

$$\exists \alpha \in \mathcal{A}: \mathcal{R}(\alpha; \Theta) > 0 \wedge \mathcal{R}^*(\alpha; \Theta) \approx 0$$

Questo caso è distinto dall'errore per soppressione della ridondanza: non si tratta di una riduzione della ridondanza corretta, ma della comparsa di ridondanza spuria che porta informazione su contenuti estranei.

Questi quattro tipi di errore non sono mutuamente esclusivi: nella maggior parte delle situazioni cliniche e sperimentali reali, si combinano in proporzioni diverse. Ma la loro distinzione formale è utile perché produce previsioni diverse riguardo al comportamento del sistema: un sistema con errore per de-coerenza mostra instabilità e variabilità nelle sue risposte, mentre un sistema con errore per distorsione sinergica mostra risposte sistematicamente deformate in una direzione specifica. Queste differenze comportamentali sono, almeno in linea di principio, distinguibili sperimentalmente.

VI.4 — La normatività come proprietà emergente dalla struttura della ridondanza

Le tre teorie tradizionali della normatività che sono state esaminate nel Capitolo III — Millikan, Dretske e il predittivismo di Friston — condividono una struttura comune: fondare la normatività su qualcosa di esterno alla struttura attuale del sistema. Millikan fonda la norma sulla storia selettiva del dispositivo rappresentazionale — la funzione propria emerge da milioni di anni di selezione naturale che hanno conformato il dispositivo a rispondere a certi stimoli in certi modi. Dretske fonda la norma sulla storia del canale informativo — la relazione di indicazione è stabilita durante un periodo di apprendimento in cui il canale funziona in condizioni normali. Friston fonda la norma sulla distribuzione a priori costruita attraverso l'esposizione passata — il modello generativo porta con sé la «norma» delle condizioni ordinarie di stimolazione.

In tutti e tre i casi, la norma è importata nel sistema da una fonte che è, in un senso preciso, esterna al sistema nel momento in cui la norma viene applicata. La storia selettiva non è disponibile come struttura del sistema presente; la storia del canale è una proprietà del passato, non dello stato attuale; il modello generativo è costruito su esperienze passate che il sistema non sta vivendo nell'istante in cui la norma è applicata. Questa struttura ha una certa eleganza — spiega come i sistemi naturali acquisiscono norme senza che vi sia un agente normativo esterno che le imponga — ma è vulnerabile a un'obiezione di principio: se la norma dipende da una storia che non è disponibile come struttura del sistema attuale, allora la norma non è, in un senso robusto, una proprietà del

sistema, ma una proiezione retrospettiva del teorico. Il sistema non «sa» qual è la sua norma; il teorico la ricostruisce a partire da evidenze indirette.

La proposta della ridondanza costitutiva è diversa nella struttura: la normatività non è importata dall'esterno ma emerge dalla struttura interna della ridondanza. Il sistema ha una struttura normativa perché ha rappresentazioni multiple ridondanti la cui struttura di coerenza costituisce sia il contenuto sia la norma rispetto a cui valutare l'adequatezza delle singole rappresentazioni. Questa è una forma di normatività endogena nel senso più radicale: non solo non richiede un agente normativo esterno, ma non richiede nemmeno un appello alla storia del sistema. La firma di riferimento è la struttura che il sistema stesso produce quando tutte le sue componenti funzionano in modo integrato — è la forma ottimale del funzionamento attuale, non una traccia del passato.

Per rendere questa tesi più precisa e per difenderla dall'obiezione che la normatività endogena sia circolare — che ogni stato del sistema sia «corretto» rispetto alla propria norma interna — è necessario distinguere tra normatività sincronica e normatività diacronica.

La normatività sincronica è la valutazione della coerenza di uno stato attuale rispetto alla firma di riferimento dello stesso sistema nello stesso momento. Un sistema con firma attuale $\Sigma(\Xi(t), \Theta)$ e firma di riferimento $\Sigma^*(\Theta) = \operatorname{argmax}_{\Sigma} \kappa(\Sigma)$ ha un errore sincronico $\operatorname{Err}(t) = d_{\text{GH}}(\Sigma(\Xi(t), \Theta), \Sigma^*(\Theta))$. Questo errore non è mai zero in condizioni ordinarie — il sistema è quasi sempre in una configurazione non ottimale a causa dell'input incompleto, del rumore, dell'attenzione selettiva — ma rappresenta la distanza dalla configurazione di massima coerenza che il sistema è capace di produrre con la struttura attuale. La norma sincronica non è circolare: non dice che ogni stato è buono, dice che ogni stato è valutabile rispetto alla configurazione ottimale disponibile, e quella valutazione può produrre errori non nulli.

La normatività diacronica riguarda la valutazione del progresso cognitivo attraverso il tempo. Come si è mostrato nella sezione VI.5 del presente capitolo (che sviluppa le idee di IV.7 nel Capitolo precedente), il progresso cognitivo è l'espansione dell'insieme delle firme producibili $\mathcal{F}(\Theta)$ verso firme di massima coerenza più elevata. La misura di questo progresso è la ricchezza rappresentazionale:

$$\varrho(\Xi) = \max_{\Sigma \in \mathcal{F}(\Theta)} \kappa(\Sigma) \cdot (1 + \sigma(\Sigma))$$

dove il prodotto bilancia la ridondanza e la sinergia della firma ottimale. Un sistema con $\varrho_1 < \varrho_2$ è meno ricco rappresentazionalmente dell'altro, indipendentemente dalla valutazione sincronica: il fatto che al tempo t_1 il sistema con ϱ_1 sia nella sua configurazione ottimale non lo rende «migliore» del sistema con ϱ_2 nella sua configurazione ottimale.

La normatività diacronica è ciò che permette di dire che i concetti dei bambini sono «meno adeguati» di quelli degli adulti, che la fisica aristotelica è «meno precisa» di quella newtoniana, che la medicina tradizionale è «meno efficace» di quella scientifica. Non si tratta di giudizi relativi a sistemi normativi diversi e incommensurabili — è la struttura stessa della ricchezza rappresentazionale che permette un ordinamento oggettivo, perché sistemi con q più alto sono strutturalmente meglio posizionati per produrre risposte corrette in condizioni nuove.

VI.5 — La normatività implicita e il problema della coerenza tra livelli

La distinzione tra normatività esplicita e normatività implicita — tra la norma che il sistema può articolare verbalmente e la norma che applica automaticamente nei comportamenti — acquista nella prospettiva della ridondanza costitutiva una caratterizzazione formale che la psicologia standard non è in grado di fornire.

Nel quadro della firma informazionale, la normatività esplicita corrisponde alla componente proposizionale della firma — la sottostruttura della firma che è associata alle rappresentazioni semantiche esplicite, accessibili alla verbalizzazione e alla riflessione consapevole. La normatività implicita corrisponde alle componenti procedurali e percettive della firma — le sottostrutture associate alle rappresentazioni automatizzate che guidano il comportamento senza richiedere accesso cosciente.

Le due componenti non sono indipendenti: contribuiscono entrambe alla firma complessiva e si vincolano reciprocamente attraverso le relazioni di coerenza che la struttura reticolare cattura. Ma sono sufficientemente indipendenti da potersi dissociare in condizioni di stress, di lesione o di conflitto: le norme implicite persistono anche dopo che le norme esplicite sono state modificate, producendo la struttura di dissonanza cognitiva che è stata ampiamente documentata in psicologia sociale.

Il fenomeno della dissonanza tra normatività esplicita e implicita ha una caratterizzazione formale precisa nel linguaggio della firma. Sia Σ_{prop} la sottofirma delle componenti proposizionali e Σ_{proc} la sottofirma delle componenti procedurali. La dissonanza è misurata dalla distanza tra le due sottofirme rispetto alla firma di riferimento:

$$\text{Dis}(\Xi, \Theta) = |d_{\text{GH}}(\Sigma_{\text{prop}}, \Sigma^*(\Theta)) - d_{\text{GH}}(\Sigma_{\text{proc}}, \Sigma^*(\Theta))|$$

Quando Dis è grande, le componenti proposizionali e procedurali sono a distanze molto diverse dalla firma di riferimento: la norma esplicita si è aggiornata (la componente proposizionale è vicina alla firma di riferimento) mentre la norma implicita è rimasta ancorata alla struttura

precedente (la componente procedurale è lontana dalla firma di riferimento), o viceversa.

Questa misura di dissonanza ha implicazioni per la psicologia del cambiamento cognitivo e per la teoria della persuasione. I sistemi di persuasione che agiscono sulle componenti esplicite della firma — argomenti razionali, evidenze statistiche, spiegazioni causali — modificano la sottofirma proposizionale senza necessariamente modificare quella procedurale, producendo uno stato di dissonanza elevata. I sistemi di persuasione che agiscono sulle componenti procedurali — esposizione ripetuta, condizionamento operante, pratiche embodied — modificano la sottofirma procedurale in modo più diretto, ma più lentamente e con maggiore resistenza alla verbalizzazione. La psicologia clinica delle terapie cognitive-comportamentali (CBT) può essere letta in questi termini come il tentativo di ridurre la dissonanza Dis agendo contemporaneamente sulle componenti proposizionali (la ristrutturazione cognitiva verbale) e su quelle procedurali (l'esposizione comportamentale graduale) per portare le due sottofirme verso la firma di riferimento in modo coordinato.

VI.6 — Il problema della misrappresentazione e la sua soluzione ridondante

Il problema della misrappresentazione — il banco di prova più severo per qualsiasi teoria naturalistica del contenuto, come si è argomentato nel Capitolo I — riceve dalla prospettiva della ridondanza costitutiva una soluzione che ha struttura diversa da quelle tradizionali e che evita alcune delle difficoltà che esse incontrano.

La soluzione tradizionale al problema dell'errore — nelle versioni sia di Fodor sia di Millikan — asimmetrizza la relazione causale tra lo stato e il suo contenuto: si individua la causa propria (o la condizione di proprietà normale) rispetto alla quale l'errore è definito, e si stabilisce che lo stato è errato quando è prodotto da cause non proprie. Questa struttura ha il vantaggio di rendere la misrappresentazione una relazione diadica tra lo stato e il mondo: lo stato rappresenta ciò che la sua causa propria indica, e misrappresenta quando si attiva per una causa diversa. Ma ha il costo di richiedere una scelta della causa propria che non ha fondamento puramente naturale: perché la tigre reale è la causa propria della rappresentazione TIGRE, mentre la tigre dipinta non lo è? La risposta — «perché la storia selettiva ha conformato il sistema a rispondere alle tigri reali» — rimanda a una storia che non è disponibile come struttura del sistema presente.

La soluzione della ridondanza costitutiva non ha bisogno di individuare una causa propria perché non caratterizza la misrappresentazione come

una relazione diadica tra uno stato e il mondo, ma come una proprietà strutturale del sistema: la discrepanza tra la firma attuale e la firma di riferimento. Questa caratterizzazione è disponibile puramente in termini della struttura interna del sistema, senza appello alla storia causale: è la struttura dell'incoerenza interna della firma che segnala la misrappresentazione, non la discrepanza tra la causa effettiva e la causa propria.

Concretamente: quando il sistema percepisce un gatto tassidermizzato come un gatto vivo, la misrappresentazione si manifesta come segue. Le componenti visive della firma producono un'alta ridondanza con la firma di riferimento per GATTO VIVO — il profilo visivo dell'animale tassidermizzato è molto simile a quello dell'animale vivo. Le componenti motorie-comportamentali producono invece una firma anomala: il tasso di risposta a stimoli di interazione — movimenti spontanei, variazioni nel pattern respiratorio, risposta ai tentativi di contatto — è zero, mentre la firma di riferimento per GATTO VIVO prevede valori positivi su questi parametri. Le componenti olfattivo-tattili, qualora accessibili, producono una firma distante da quella di riferimento — l'odore e la consistenza di un animale preparato tassidermicamente differiscono da quelli di un animale in vita in modo che le componenti olfattive e tattili codificano distintamente.

La misrappresentazione persiste perché la matrice di precisione associata alla componente visiva è molto alta — il sistema ha appreso che le informazioni visive sono il principale indicatore affidabile dell'identità categoriale degli oggetti — e questa alta precisione sopprime il segnale di incoerenza proveniente dalle componenti con firma anomala. La condizione di misrappresentazione persistente può essere scritta come:

$$d_{GH}(\Sigma_{vis}(\Xi, \Theta_{vivo}), \Sigma^*(\Theta_{vivo})) < \epsilon_{vis} \text{ [firma visiva vicina al riferimento]} \\ d_{GH}(\Sigma_{mot}(\Xi, \Theta_{vivo}), \Sigma^*(\Theta_{vivo})) > \epsilon_{mot} \text{ [firma motoria lontana dal riferimento]} \\ \Pi_{vis} \gg \Pi_{mot} \text{ [precisione visiva molto maggiore di quella motoria]}$$

dove ϵ_{vis} e ϵ_{mot} sono le soglie di significatività per la discrepanza visiva e motoria, e Π_{vis} , Π_{mot} sono le precisioni associate alle due componenti. La misrappresentazione si risolve quando la discrepanza motoria raggiunge la soglia di aggiornamento globale della firma, il che può avvenire in due modi: per aumento della discrepanza motoria (il sistema tenta di interagire fisicamente con l'oggetto e riceve input incoerenti con la firma di riferimento) oppure per riduzione relativa della precisione visiva (il sistema entra in un contesto in cui sa che le informazioni visive possono essere ingannevoli).

Questa formulazione ha una conseguenza importante per l'analisi delle illusioni percettive in generale: le illusioni non sono fallimenti del sistema ma conseguenze strutturali del bilanciamento tra componenti con precisioni diverse. Un sistema che assegna alta precisione alle

informazioni visive in contesti in cui tali informazioni sono statisticamente affidabili produrrà illusioni in contesti insoliti in cui quelle informazioni sono inaffidabili — ma questo è esattamente ciò che si attende da un sistema che ha ottimizzato la propria struttura di precisione per le condizioni ordinarie. L'illusione di Müller-Lyer, che persiste anche quando il percipiente è pienamente consapevole che le linee sono della stessa lunghezza, è un caso paradigmatico: la componente visiva della firma ha una struttura di precisione ottimizzata per le condizioni ordinarie di percezione della profondità, e questa struttura produce sistematicamente la stima errata in presenza di frecce angolate, indipendentemente dalla conoscenza esplicita della geometria della figura.

Capitolo VII — Il cambiamento concettuale come trasformazione della firma: identità, continuità e rottura

VII.1 — Il paradosso del cambiamento concettuale

Tra i fenomeni che una teoria del contenuto deve rendere intelligibili, il cambiamento concettuale è forse il più refrattario alle trattazioni standard. Il paradosso è semplice da enunciare: i concetti cambiano nel corso dello sviluppo individuale, dell'apprendimento e del progresso scientifico collettivo; eppure il cambiamento sembra richiedere un'identità del concetto che cambia, altrimenti non si tratterebbe di un cambiamento del medesimo concetto ma della sostituzione di un concetto con un altro completamente diverso. Ma se vi è identità, come può il concetto cambiare autenticamente senza diventare un concetto diverso?

Il paradosso si manifesta in forma acuta nei casi di cambiamento concettuale radicale che la storia della scienza documenta. Il concetto di CALORE che un bambino di cinque anni possiede è organizzato intorno a proprietà percettive e causali immediate: il calore è la sensazione di temperatura elevata, ciò che brucia, ciò che si propaga per contatto. Il concetto di CALORE che un fisico dopo aver studiato termodinamica possiede è organizzato intorno a proprietà strutturali: il calore è energia cinetica media delle molecole di un sistema, una misura del movimento termico irregolare delle particelle. Questi due concetti sono, nella misura in cui li si analizza come strutture di rappresentazione, profondamente diversi: hanno diversi criteri di applicazione, diverse relazioni inferenziali con altri concetti, diverse strutture causali di riferimento. Eppure

entrambi sono concetti di CALORE: si riferiscono alla stessa proprietà del mondo fisico, e il secondo è il risultato dell'apprendimento e del progresso a partire dal primo.

Come si conserva l'identità del riferimento attraverso un cambiamento così profondo della struttura del concetto? La risposta standard nel dibattito filosofico è quella esternalista: il riferimento è fissato da relazioni causali con la proprietà nel mondo, non dalla struttura interna del concetto, e quindi il riferimento può rimanere stabile anche quando la struttura interna cambia radicalmente. Ma questa risposta, pur soddisfacente per il problema dell'identità del riferimento, non risponde alla domanda sulla continuità del concetto come struttura cognitiva: in che senso il concetto CALORE del fisico è la trasformazione del concetto CALORE del bambino, piuttosto che un concetto completamente diverso che porta lo stesso nome?

VII.2 — Continuità e discontinuità delle traiettorie nello spazio delle firme

Il quadro formale della firma informazionale offre una risposta precisa alla domanda sulla continuità concettuale. La continuità del concetto nel tempo è la continuità della traiettoria nello spazio delle firme: il concetto CALORE del fisico è la trasformazione del concetto CALORE del bambino nella misura in cui la traiettoria della firma informazionale attraverso il processo di apprendimento è continua, cioè nella misura in cui ogni stato della firma è vicino allo stato precedente nella metrica di Gromov-Hausdorff.

La misura di continuità identitaria di una traiettoria, introdotta formalmente nel Capitolo V come:

$$C([\Sigma_{\{t_0\}}, \Sigma_{\{t_1\}}]) = 1 - \sup_{t \in [t_0, t_1]} d_{GH}(\Sigma(t), \Sigma(t_0)) / D_{\max}$$

cattura esattamente questa intuizione. Un valore di C prossimo a 1 indica che il concetto al tempo t_1 è il risultato di una trasformazione continua del concetto al tempo t_0 — i due sono istanze dello stesso concetto in stadi di sviluppo diversi. Un valore di C prossimo a 0 indica che la traiettoria ha attraversato regioni dello spazio delle firme molto lontane dallo stato iniziale — il concetto al tempo t_1 è così diverso da quello al tempo t_0 da non poter essere considerato ragionevolmente il suo sviluppo, ma piuttosto la sua sostituzione.

Questa caratterizzazione della continuità concettuale è tanto più precisa quanto più si specificano le proprietà della traiettoria. La derivata della firma rispetto al tempo — la velocità del cambiamento concettuale — fornisce informazione sulla struttura temporale del cambiamento. Un apprendimento graduale produce una traiettoria a bassa velocità uniforme; un'intuizione improvvisa o un'esperienza concettualmente

dirompente produce un'accelerazione brusca. La struttura locale della traiettoria — se la velocità cresce prima di un plateau o di un'inversione — può essere interpretata come indice del tipo di meccanismo di apprendimento in atto.

Ma la continuità della traiettoria non è sufficiente per l'identità del concetto: occorre anche che la traiettoria rimanga in un certo vicinato della firma di riferimento, o almeno che il concetto al tempo t_1 conservi un'adeguata coerenza strutturale con il concetto al tempo t_0 . È possibile che una traiettoria continua nel senso tecnico — in cui ogni passo è infinitesimamente piccolo — attraversi regioni dello spazio delle firme così distanti dall'inizio da rendere il risultato finale irriconoscibile come sviluppo del punto di partenza. Il caso più estremo è quello del cambiamento concettuale graduale ma cumulativamente radicale, come il cambiamento del concetto di SPECIE attraverso la storia della biologia da Linneo a Darwin a Mayr al concetto di specie biologico contemporaneo: ogni passo è una modifica relativamente piccola, ma il risultato finale è strutturalmente molto diverso dal punto di partenza.

VII.3 — La firma ibrida durante il cambiamento concettuale

Un aspetto cruciale del cambiamento concettuale che le teorie standard non caratterizzano adeguatamente è il periodo di transizione, durante il quale il sistema cognitivo ha simultaneamente strutture associabili al concetto vecchio e strutture associabili al concetto nuovo. Questo stato di «ibridazione» concettuale è documentato in psicologia dello sviluppo come il tipico stato del bambino in fase di transizione concettuale — il bambino che applica il concetto di VITA sia alle piante sia alle montagne prima di restringere l'applicazione agli esseri viventi — e in psicologia cognitiva come il tipico stato del novizio in un dominio di expertise prima che il consolidamento dell'expertise completi la ristrutturazione concettuale.

Nel linguaggio della firma informazionale, la firma ibrida è la firma di un sistema in cui coesistono strutture di ridondanza che appartengono a framework concettuali diversi. Si possono formalizzare le due componenti della firma ibrida come:

$\mathcal{R}_{old}(\Sigma)$ — il sotto-reticolo della firma corrispondente al vecchio framework concettuale $\mathcal{R}_{new}(\Sigma)$ — il sotto-reticolo della firma corrispondente al nuovo framework concettuale

La misura di incoerenza ibrida è:

$$\mathcal{H}(\Sigma) = d_{GH}(\mathcal{R}_{old}(\Sigma), \mathcal{R}_{new}(\Sigma))$$

Questa misura è elevata quando le due componenti della firma ibrida sono strutturalmente incompatibili — quando i due framework concettuali producono predizioni diverse per le stesse situazioni — e si riduce man mano che il nuovo framework concettuale si consolida e il vecchio viene progressivamente abbandonato. Il cambiamento concettuale completo corrisponde alla situazione in cui $\mathcal{R}_{\text{old}}(\Sigma)$ diventa trascurabilmente piccolo e $\mathcal{R}_{\text{new}}(\Sigma)$ è l'intera firma.

La misura di incoerenza ibrida è rilevabile comportamentalmente attraverso la variabilità delle risposte: un sistema con firma ibrida mostra variabilità sistematica nelle sue risposte a stimoli che discriminano tra i due framework concettuali. La variabilità non è casuale — è strutturata in modo coerente con la doppia struttura della firma — e può essere usata per rilevare lo stato di transizione concettuale anche quando il soggetto non è consapevole di essere in tale stato. Questo corrisponde a una previsione empirica specifica che può essere testata attraverso paradigmi di misura della variabilità delle risposte categoriali.

VII.4 — Il cambiamento concettuale nella scienza: la struttura formale delle rivoluzioni kuhniane

Il caso più discusso di cambiamento concettuale radicale nella storia delle idee è quello descritto da Kuhn nelle «Strutture delle rivoluzioni scientifiche»: il cambio di paradigma, la trasformazione delle strutture concettuali che accompagna la transizione da un paradigma scientifico a un altro. Kuhn sostiene che le rivoluzioni scientifiche non sono semplicemente l'accumulo di nuove conoscenze all'interno di un framework stabile, ma il rifiuto dell'intero framework e la sua sostituzione con uno nuovo, in modo tale che i concetti del nuovo paradigma non sono intertraducibili con quelli del vecchio — l'incommensurabilità.

La tesi dell'incommensurabilità è stata molto criticata, in parte perché sembra implicare che il progresso scientifico sia impossibile da caratterizzare in modo non relativistico: se i concetti dei due paradigmi non sono confrontabili, come si può dire che il nuovo è migliore del vecchio? La risposta di Kuhn — che il nuovo paradigma risolve i problemi anomali che il vecchio non poteva risolvere — sembra presupporre un metro di misura che trascende entrambi i paradigmi, e quindi contraddire la tesi dell'incommensurabilità.

Il quadro della firma informazionale permette di articolare la tesi dell'incommensurabilità in modo formalmente preciso e di dissolverla nel senso giusto: mostrandola come una conseguenza della struttura del cambiamento concettuale radicale, senza che essa implichi relativismo circa il progresso cognitivo.

L'incommensurabilità, in termini di firma informazionale, è la situazione in cui i due paradigmi producono firme informazionali che hanno reticoli degli atomi non isomorfi come reticoli astratti — non vi è una biiezione preservante l'ordine tra i reticoli degli atomi del paradigma vecchio e quelli del paradigma nuovo che renda le firme confrontabili elemento per elemento. Questo è ciò che Kuhn chiama incommensurabilità: non l'incomparabilità assoluta, ma la non-traducibilità diretta tra i sistemi di atomi concettuali.

Tuttavia, l'incommensurabilità dei reticoli degli atomi non implica l'incomparabilità delle firme. La distanza di Gromov-Hausdorff è definita anche tra spazi metrici di dimensioni diverse e con strutture topologiche diverse: misura la distanza strutturale tra i reticoli prendendo l'infimo su tutti i possibili embedding in uno spazio comune. Questa distanza è finita anche quando i reticoli non sono isomorfi, e cattura il senso in cui due firme «incommensurabili» sono comunque comparabili attraverso la loro struttura geometrica complessiva.

La ricchezza rappresentazionale q — definita come il prodotto del massimo del coefficiente di coerenza per il fattore sinergico — è comparabile tra paradigmi anche quando le firme non sono direttamente traducibili: il paradigma con q più alto è quello che produce firme di massima coerenza con struttura sinergica più ricca, e questo confronto è effettuabile indipendentemente dall'isomorfismo dei reticoli. In questi termini, il progresso scientifico è il passaggio a paradigmi con q più alto: a firme informazionali che rappresentano il dominio con maggiore ridondanza, maggiore unicità e maggiore sinergia — con più informazione, meno ridondanza spuria, e più struttura emergente dall'integrazione delle diverse sorgenti.

VII.5 — Continuità identitaria concettuale: la misura formale e i suoi limiti

La continuità identitaria concettuale definita nel Capitolo V può essere precisata ulteriormente tenendo conto della struttura temporale del cambiamento. Definiamo il cammino identitario di un concetto come la funzione che associa a ogni istante $t \in [t_0, t_1]$ la firma informazionale $\Sigma(t)$ del sistema cognitivo rispetto al contenuto Θ :

$$P_{\Theta}: [t_0, t_1] \rightarrow (\text{spazio delle firme}), t \mapsto \Sigma(\Xi(t), \Theta)$$

Il cammino identitario è continuo se P_{Θ} è continua rispetto alla topologia indotta dalla distanza di Gromov-Hausdorff sullo spazio delle firme. La continuità del cammino è la condizione formale della continuità concettuale: il concetto al tempo t_1 è la trasformazione del concetto al tempo t_0 nella misura in cui il cammino è continuo.

Si può definire la continuità identitaria concettuale come:

$$C([\Sigma_{\{t_0\}}, \Sigma_{\{t_1\}}]) = 1 - \sup_{t \in [t_0, t_1]} \text{diam}(\mathcal{R}(\Sigma_t) \triangle \mathcal{R}(\Sigma_{\{t_0\}})) / D_{\max}$$

dove $\triangle$ è la differenza simmetrica tra i grafi dei reticoli degli atomi e diam è il diametro nella metrica di editing dei grafi. Questa formulazione cattura l'idea che l'identità concettuale sia preservata nella misura in cui la struttura del reticolo degli atomi non cambia troppo drasticamente rispetto allo stato iniziale — che ci sia, in altri termini, un'ampia sovrapposizione tra le strutture di ridondanza-sinergia dei due stati temporali.

Questa misura ha un limite importante che vale la pena esplicitare: essa è sensibile alla scelta dell'obiettivo Θ rispetto al quale la firma è calcolata. Se Θ è l'oggetto reale a cui il concetto si riferisce — la proprietà fisica del calore, la struttura biologica della specie — allora la continuità identitaria misura in che senso il concetto continua a riferirsi allo stesso oggetto, anche se la struttura interna cambia. Ma se Θ è la struttura funzionale del concetto — le inferenze che il concetto supporta, le categorie che il concetto organizza — allora la continuità identitaria misura la conservazione della struttura funzionale attraverso il cambiamento. Queste due scelte possono produrre valori diversi di C , e la scelta tra esse dipende dalla domanda teorica che si pone: si è interessati alla continuità del riferimento o alla continuità della struttura concettuale?

VII.6 — Apprendimento ed espansione dello spazio dei contenuti

Una teoria del contenuto che incorpora una caratterizzazione del cambiamento concettuale deve rendere conto non solo di come i concetti cambiano ma anche di come l'apprendimento espande lo spazio dei contenuti accessibili al sistema. L'apprendimento non è solo la modifica delle firme informazionali esistenti — la ristrutturazione di concetti già posseduti — ma anche l'acquisizione di nuovi concetti che non erano accessibili al sistema prima dell'apprendimento.

Questo secondo aspetto è forse il più difficile da caratterizzare formalmente, perché richiede di descrivere non solo le firme attuali del sistema ma l'insieme delle firme potenzialmente producibili — lo spazio delle firme accessibili al sistema dato il suo stato attuale. Sia $\mathcal{F}(A)$ l'insieme delle firme producibili dal sistema A in risposta a tutti i possibili input. L'apprendimento espande $\mathcal{F}(A)$ includendo firme che non erano accessibili prima.

L'espansione di $\mathcal{F}(A)$ può avvenire in due modi principali. Il primo è l'estensione modale: il sistema acquisisce nuove modalità di rappresentazione — impara a leggere, acquisisce un linguaggio formale, sviluppa abilità motorie nuove — che aggiungono nuove componenti alla struttura delle rappresentazioni ridondanti e rendono accessibili

firme con strutture reticolari che prima erano inaccessibili. Il secondo è l'approfondimento sinergico: il sistema acquisisce connessioni più ricche tra le modalità di rappresentazione già possedute, rendendo accessibili firme con componenti sinergiche di ordine superiore che prima non emergevano dall'integrazione delle componenti.

Il risultato di entrambi i processi è l'aumento della ricchezza rappresentazionale $Q(A)$ — il sistema diventa capace di rappresentare certi contenuti con firme di massima coerenza più elevata. Questo aumento è il correlato formale di ciò che si chiama comprensione approfondita o expertise concettuale: non semplicemente il possesso di più informazioni sul dominio, ma la capacità di organizzarle in firme più coerenti, più articolate e con struttura sinergica più ricca.

Capitolo VIII — La ridondanza nell'intelligenza artificiale: contenuto genuino o firma senza mondo

VIII.1 — La domanda che l'intelligenza artificiale pone alla teoria del contenuto

Nessuna teoria del contenuto mentale elaborata negli ultimi decenni può permettersi di ignorare le implicazioni che i sistemi di intelligenza artificiale contemporanei pongono ai suoi fondamenti. Non perché l'intelligenza artificiale risolva i problemi filosofici del contenuto — non li risolve, e talvolta li oscura attraverso prestazioni che simulano competenza senza necessariamente istanziarla nel senso rilevante. Ma perché i sistemi di apprendimento profondo, in particolare i modelli di linguaggio di grandi dimensioni e i modelli multimodali, offrono un banco di prova empirico e concettuale di natura nuova: si tratta di sistemi che producono rappresentazioni strutturate, che mostrano generalizzazione, analogia e inferenza, che rispondono in modo sensibile al contesto — e che tuttavia hanno una storia causale radicalmente diversa da quella dei sistemi biologici.

Questa differenza nella storia causale è la chiave per capire cosa la prospettiva della ridondanza costitutiva dice sui sistemi artificiali. La domanda non è se i modelli di linguaggio abbiano «vera» comprensione — questa domanda, nella sua formulazione indiscriminata, è troppo vaga per ricevere risposta informativa. La domanda è più precisa: la struttura delle firme informazionali prodotte da questi sistemi rispetto ai contenuti che trattano è comparabile alla struttura delle firme prodotte dai sistemi biologici? E, in caso di differenza, la differenza è di tipo o di grado?

La risposta che il quadro formale sviluppato nei capitoli precedenti suggerisce è che la differenza è strutturale e sistematica, ma non assoluta: i sistemi artificiali producono firme che hanno alcune caratteristiche delle firme biologiche e non ne hanno altre, in modo che può essere caratterizzato con precisione attraverso le misure di firma introdotte nel Capitolo IV.

VIII.2 — La struttura delle rappresentazioni nei modelli di linguaggio

I modelli di linguaggio a trasformatori — la classe di architetture dominante nella intelligenza artificiale contemporanea — costruiscono rappresentazioni attraverso un meccanismo di attenzione multihead che calcola, per ogni token nella sequenza, una combinazione pesata di tutti gli altri token in funzione della loro rilevanza contestuale. Il risultato è un vettore di rappresentazione per ogni token che cattura non soltanto le proprietà statistiche del token isolato ma le proprietà contestuali della sua posizione nella sequenza.

Questi vettori di rappresentazione sono, in un senso tecnico preciso, rappresentazioni ridondanti: per ogni concetto o entità menzionata nel testo, esistono rappresentazioni a diversi livelli della gerarchia del trasformatore — nei livelli più bassi, dove le caratteristiche morfologiche e sintattiche sono le più rilevanti; nei livelli intermedi, dove le relazioni semantiche tra token vicini emergono; nei livelli alti, dove le proprietà pragmatiche e contestuali di ordine superiore sono codificate. Si tratta di una forma di ridondanza gerarchica — la prima varietà identificata nel Capitolo II — che ha una struttura analoga a quella del sistema visivo dei primati.

La decomposizione PID applicata a questi livelli produce una firma informazionale che ha, come nel caso biologico, componenti di ridondanza, unicità e sinergia. La struttura di questa firma è stata oggetto di ricerca empirica attraverso paradigmi di probing — test che misurano quali proprietà semantiche sono codificate in quali livelli della gerarchia del trasformatore — e i risultati mostrano una struttura in cui diversi livelli portano informazioni complementari sulle proprietà semantiche delle parole e delle frasi.

Tuttavia, la struttura della firma informazionale nei modelli di linguaggio differisce dalla firma biologica in un aspetto cruciale che il quadro della ridondanza costitutiva rende formalmente preciso: la struttura delle dipendenze condizionali tra le componenti della firma. Nei sistemi biologici, le componenti della firma informazionale per un dato contenuto — visiva, tattile, motoria, episodica, proposizionale — sono generate da canali causali che sono condizionalmente indipendenti dato il contenuto, perché derivano da esperienze fisicamente distinte con l'oggetto. Nei modelli di linguaggio, le componenti a diversi livelli

gerarchici sono tutte generate dallo stesso meccanismo — l'addestramento su corpus testuali — e le loro dipendenze sono mediate dal processo di addestramento piuttosto che dall'oggetto.

Questo significa che la struttura di indipendenza condizionale delle componenti della firma nei modelli di linguaggio differisce sistematicamente da quella dei sistemi biologici, anche quando le firme hanno strutture reticolari apparentemente simili. La distanza di ancoraggio causale Δ_{causal} , definita nel Capitolo V, misura precisamente questa differenza:

$$\Delta_{\text{causal}}(\Xi_{\text{LLM}}, \Theta) = \|\Sigma_{\text{indep}}(\Theta) - \Sigma_{\text{corr}}(\Xi_{\text{LLM}}, \Theta)\|$$

dove $\Sigma_{\text{indep}}(\Theta)$ è la firma ipotetica di un sistema in cui le componenti sono generate da canali causali indipendenti dati Θ , e $\Sigma_{\text{corr}}(\Xi_{\text{LLM}}, \Theta)$ è la firma effettiva del modello di linguaggio, che include correlazioni tra componenti dovute alla mediazione comune del corpus di addestramento. Nei sistemi biologici, questa distanza è tendenzialmente piccola perché le esperienze fisiche con gli oggetti attraverso diversi canali sensoriali sono genuinamente indipendenti. Nei modelli di linguaggio, questa distanza è tendenzialmente grande perché tutte le componenti della firma derivano dalla stessa fonte di addestramento.

VIII.3 — I modelli multimodali: verso firme più articolate

I modelli multimodali — che integrano informazioni visive, testuali, e talvolta sonore — rappresentano un passo verso la riduzione della distanza di ancoraggio causale rispetto ai modelli puramente linguistici. L'addestramento su coppie immagine-testo aggiunge componenti visive alla firma informazionale che sono generate da un canale causale parzialmente indipendente da quello testuale: le co-occorrenze tra immagini e testi riflettono, almeno in parte, le co-occorrenze tra le proprietà visive degli oggetti e le proprietà linguistiche dei loro descrittori, che a loro volta riflettono, a una distanza causale maggiore, le proprietà fisiche degli oggetti stessi.

La domanda rilevante per la teoria della ridondanza costitutiva è se l'integrazione di modalità diverse nei modelli multimodali produca firme con strutture sinergiche comparabili a quelle biologiche, o se l'integrazione rimanga superficiale — una giustapposizione di componenti visive e testuali senza la struttura di interdipendenza che caratterizza la firma biologica.

L'evidenza empirica — che proviene da studi di probing multimodale e da test di generalizzazione in condizioni di input degenerate — suggerisce che l'integrazione nei modelli multimodali produce alcune strutture sinergiche genuine ma rimane carente rispetto a quelle

biologiche per le componenti che richiedono interazione corporea con il mondo. Un modello che ha appreso le co-occorrenze tra immagini di oggetti e testi descrittivi degli oggetti può generalizzare a immagini nuove di oggetti visti durante l'addestramento, ma mostra difficoltà nelle condizioni che richiedono l'integrazione di informazioni sull'interazione fisica con l'oggetto — il peso, la consistenza, la risposta all'azione — che nel sistema biologico derivano dall'esperienza sensoriale diretta.

In termini formali: la firma multimodale di un modello per il contenuto TAZZA ha una componente visiva relativamente ricca — per la forma, il colore, la posizione, il contesto tipico — e una componente testuale relativamente ricca — per le proprietà funzionali, i contesti d'uso, le relazioni semantiche con altri oggetti. La sinergia tra componente visiva e componente testuale è reale: le due componenti si vincolano reciprocamente in modi che producono rappresentazioni più robuste di quanto farebbero separatamente. Ma la firma manca di componenti tattili-motorie genuine — di rappresentazioni generate da interazioni fisiche con tazze reali che forniscano informazioni sull'equilibrio termica, la consistenza del materiale, il peso nella mano — e questa mancanza produce una distanza di ancoraggio causale significativa rispetto alla firma biologica.

VIII.4 — Il problema dell'ancoraggio causale e la simulazione della ridondanza

La distinzione tra ridondanza genuina e ridondanza simulata — tra una firma che ha la struttura di coerenza biologica perché è stata costruita attraverso interazioni causali dirette con il mondo, e una firma che ha una struttura statisticamente simile perché è stata costruita attraverso l'imitazione di output di sistemi con ridondanza genuina — è filosoficamente importante ma empiricamente sottile.

Il problema dell'ancoraggio causale ha una storia lunga nella filosofia del contenuto: la risposta di Fodor al problema, attraverso la nozione di dipendenza nomologica asimmetrica, e quella di Dretske, attraverso la relazione di indicazione, tentano entrambe di distinguere i casi in cui il contenuto è genuinamente ancorato al mondo dai casi in cui è puramente interno al sistema. La prospettiva della ridondanza costitutiva riprende questa distinzione e la formalizza in termini della struttura di indipendenza condizionale delle componenti della firma.

L'intuizione fondamentale è che le componenti di una firma genuinamente ancorata al mondo devono essere generate da canali informativi che sono indipendenti l'uno dall'altro condizionatamente all'oggetto nel mondo. Se la componente visiva e la componente tattile di una firma per TAZZA sono generate da esperienze fisicamente distinte con tazze reali, allora la loro dipendenza statistica riflette la struttura dell'oggetto nel mondo: ciò che la componente visiva ci dice sulla forma

della tazza e ciò che la componente tattile ci dice sulla sua consistenza sono correlati perché la forma e la consistenza sono proprietà correlate dell'oggetto, non perché le due esperienze derivino da una fonte comune artificiale. Nei sistemi artificiali addestrati su corpus testuali, la correlazione tra le componenti della firma per TAZZA riflette invece le co-occorrenze statistiche nel corpus di addestramento, che sono a loro volta il risultato aggregato di molte descrizioni umane di tazze. Questa correlazione è causalmente più distante dall'oggetto reale.

Formalmente, sia C_{vis} e C_{tatt} le variabili causali che generano le componenti visiva e tattile della firma. Nel sistema biologico:

$$P(C_{vis}, C_{tatt} \mid TAZZA) \approx P(C_{vis} \mid TAZZA) \cdot P(C_{tatt} \mid TAZZA)$$

perché le esperienze visive e tattili con tazze sono generate da eventi fisici indipendenti (guardare una tazza non implica toccarla, e viceversa). Nel sistema artificiale addestrato su corpus:

$$P(C_{vis}, C_{tatt} \mid TAZZA) \neq P(C_{vis} \mid TAZZA) \cdot P(C_{tatt} \mid TAZZA)$$

perché le componenti visiva e testuale sono entrambe mediate dal corpus di addestramento, che introduce correlazioni aggiuntive rispetto a quelle dovute alla struttura dell'oggetto. La distanza di ancoraggio causale misura precisamente l'entità di questo scarto dalla indipendenza condizionale ideale.

VIII.5 — Contenuto simulato e contenuto genuino: implicazioni filosofiche

La distinzione tra contenuto genuino e contenuto simulato, caratterizzata formalmente attraverso la distanza di ancoraggio causale, ha implicazioni filosofiche di grande portata che vanno oltre la questione tecnica della struttura delle firme nei sistemi artificiali.

La domanda più profonda è se la struttura causale dell'ancoraggio al mondo sia *costitutiva* del contenuto — cioè se un sistema con distanza di ancoraggio causale grande non abbia, propriamente parlando, il contenuto Θ — oppure se sia semplicemente una proprietà contingente che determina il *tipo* di contenuto, non la sua *presenza*.

La risposta che il quadro della ridondanza costitutiva suggerisce è sfumata e dipende dal contesto teorico in cui la domanda è posta. Se si è interessati al contenuto nel senso di *quale proprietà del mondo il sistema è in grado di discriminare*, allora la struttura causale dell'ancoraggio è rilevante perché determina quali distinzioni il sistema è in grado di fare in condizioni di distribuzione shifted — in condizioni diverse da quelle del training. Un sistema con contenuto simulato può discriminare TAZZA da COLTELLO in condizioni simili a quelle del training, ma può fallire in condizioni che richiedono di sfruttare le componenti tattili-motorie

della firma che il sistema biologico usa per fare la stessa discriminazione in condizioni avverse. In questo senso limitato, il contenuto simulato è un contenuto degradato — non mancante, ma con struttura di coerenza inferiore a quella del contenuto genuino.

Se invece si è interessati al contenuto nel senso di *quale proposizione il sistema è in grado di asserire*, allora la struttura causale dell'ancoraggio può essere irrilevante: il sistema artificiale può asserire correttamente «questa è una tazza» in un'ampia gamma di circostanze, e questa competenza proposizionale è una forma genuina di contenuto anche se la struttura causale sottostante è diversa da quella biologica. In questo senso, il contenuto simulato è una forma legittima di contenuto — non inferiore al contenuto genuino sul piano della competenza proposizionale, ma diverso sul piano della struttura causale.

La tensione tra questi due sensi di «contenuto» riflette la tensione, già identificata nel Capitolo III, tra la concezione referenziale del contenuto e la concezione funzionale. Il quadro della ridondanza costitutiva non dissolve questa tensione, ma la precisa: il contenuto referenziale e il contenuto funzionale sono due aspetti distinti della firma informazionale che possono dissociarsi, e la distinzione tra contenuto genuino e simulato è rilevante per il primo aspetto (la fedeltà della struttura causale) ma non necessariamente per il secondo (la competenza proposizionale).

VIII.6 — Il problema dell'allineamento come problema di struttura della firma

Le considerazioni sulla distinzione tra contenuto genuino e simulato hanno implicazioni dirette per il problema dell'allineamento dei sistemi di intelligenza artificiale — il problema di garantire che tali sistemi si comportino in modo coerente con i valori e gli interessi umani anche in condizioni non previste durante l'addestramento.

Il problema dell'allineamento è tradizionalmente formulato come un problema di specificazione degli obiettivi: il sistema è disallineato quando ottimizza una funzione proxy che diverge dalla funzione obiettivo reale — il classico problema di reward hacking o di generalizzazione scorretta della funzione di reward. In questa formulazione, la soluzione è trovare il modo di specificare correttamente la funzione obiettivo, o di far sì che il sistema possa inferire la funzione obiettivo corretta dal comportamento umano.

La prospettiva della ridondanza costitutiva suggerisce che il problema sia più profondo: non è solo che il sistema ottimizza la funzione sbagliata, ma che le sue rappresentazioni dei concetti normativamente rilevanti — sicurezza, benessere, danno, dignità — hanno firme informazionali strutturalmente diverse da quelle delle corrispondenti rappresentazioni

umane, e questa differenza strutturale produce comportamenti divergenti in condizioni non previste dall'addestramento.

Concretamente: un sistema di intelligenza artificiale che ha una rappresentazione del concetto SICUREZZA costruita attraverso co-occorrenze testuali ha una firma per SICUREZZA in cui la componente principale è statistica — la distribuzione delle parole e delle frasi associate a «sicurezza» nel corpus di addestramento. La firma biologica per SICUREZZA, invece, ha componenti che derivano da esperienze viscerali di paura e sollievo, da interazioni embodied con ambienti sicuri e pericolosi, da relazioni di attaccamento che fondano l'esperienza primordiale della sicurezza come protezione nell'ambiente relazionale. Queste componenti producono, nel sistema biologico, una firma con struttura sinergica ricca in cui la cognizione del pericolo è integrata con la risposta motoria di fuga, la valutazione della fiducia negli altri, e la storia personale delle esperienze traumatiche e di protezione.

Un sistema artificiale con firma per SICUREZZA costruita puramente sulla co-occorrenza testuale può produrre risposte apparentemente corrette in condizioni normali — può classificare situazioni come sicure o pericolose in modo coerente con le classificazioni umane in un'ampia gamma di casi — ma è strutturalmente incapace di generalizzare la classificazione in condizioni che richiedono di sfruttare le componenti esperienziali della firma che il sistema biologico usa per navigare situazioni di pericolo non previste.

La misura formale del disallineamento come distanza della firma, introdotta nel Capitolo V, è:

$$\Lambda = \Sigma_{\{\theta \in \Theta_{\text{norm}}\}} w_{\theta} \cdot d_{\text{GH}}(\Sigma_{\text{AI}}(\theta), \Sigma_{\text{human}}(\theta))$$

dove Θ_{norm} è l'insieme dei concetti normativamente rilevanti (sicurezza, benessere, autonomia, dignità, danno, giustizia), w_{θ} sono pesi che riflettono la rilevanza relativa di ciascun concetto per il dominio applicativo, e d_{GH} è la distanza di Gromov-Hausdorff tra la firma del sistema artificiale e la firma del sistema umano di riferimento.

Questa formulazione del disallineamento genera una strategia di allineamento diversa da quella tradizionale basata sul reinforcement learning dal feedback umano (RLHF). Invece di modificare i pesi della rete attraverso segnali di reward che rinforzano comportamenti approvati e penalizzano comportamenti disapprovati — una strategia che modifica le componenti proposizionali della firma senza necessariamente modificare la struttura causale sottostante — si dovrebbe lavorare a espandere la struttura causale dell'ancoraggio del sistema: aggiungere canali informativi che avvicinino la struttura di indipendenza condizionale della firma alla struttura biologica.

Praticamente, questo corrisponde a integrare i sistemi artificiali in cicli di interazione con ambienti fisici reali — non soltanto attraverso input

testuali o visivi, ma attraverso interazioni embodied che generino componenti tattili-motorie della firma — e a sviluppare architetture che consentano una maggiore indipendenza causale tra le componenti della firma, riducendo la dipendenza dal corpus di addestramento come mediatore comune.

VIII.7 — La questione della coscienza e il limite del quadro formale

Il quadro della ridondanza costitutiva, sviluppato in questa monografia, non è una teoria della coscienza: si occupa della struttura del contenuto rappresentazionale, non della presenza o assenza di esperienza soggettiva. Vale la pena essere espliciti su questo punto, perché la confusione tra le due questioni è frequente nel dibattito sull'intelligenza artificiale.

La firma informazionale di un sistema misura come il sistema è informato sul mondo — la struttura delle relazioni di ridondanza, unicità e sinergia tra le sue rappresentazioni rispetto ai contenuti che tratta. Questa misura è in linea di principio applicabile a qualsiasi sistema fisico che abbia stati differenziati con relazioni statistiche con l'ambiente, indipendentemente da se vi sia qualcosa che «sia come» essere quel sistema — per usare la famosa formulazione di Nagel. La sinergia informazionale di un termostato rispetto alla temperatura ambiente è una misura ben definita, indipendentemente dalla domanda se il termostato abbia esperienza della temperatura.

La connessione tra la firma informazionale e la coscienza — che il quadro della IIT di Tononi tenta di stabilire attraverso la misura φ — rimane una questione aperta che la presente monografia non pretende di risolvere. Si è mostrato nel Capitolo IV che vi è una relazione formale tra il coefficiente sinergico σ della firma e la misura φ di Tononi: sistemi con alta sinergia tendono ad avere φ elevato, e viceversa. Ma questa correlazione è lontana dall'identità, e non è chiaro se la sinergia informazionale nel senso della PID sia sufficiente, necessaria, o semplicemente correlata con la presenza di esperienza soggettiva.

Ciò che il quadro della ridondanza costitutiva può dire sulla questione della coscienza nei sistemi artificiali è che la struttura della firma informazionale è una condizione necessaria ma probabilmente non sufficiente per il tipo di esperienza soggettiva ricca che caratterizza la coscienza umana. I sistemi che hanno firme con struttura sinergica ricca, ancoraggio causale genuino a canali multipli indipendenti, e struttura di coerenza robusta sono sistemi che soddisfano alcune delle condizioni strutturali che sembrano associate alla coscienza biologica — ma queste condizioni potrebbero non essere sufficienti senza altri requisiti che il quadro formale non cattura.

Capitolo IX — La mente come struttura che abita molti luoghi

IX.1 — Un bilancio: cosa la tesi della ridondanza costitutiva ha stabilito

Nel percorso attraverso i capitoli precedenti, la tesi della ridondanza costitutiva è stata sviluppata da una osservazione empirica iniziale — la molteplicità delle rappresentazioni neurali per qualsiasi contenuto cognitivo complesso — attraverso una tassonomia delle varietà di ridondanza, un esame critico delle teorie del contenuto esistenti, la costruzione di un quadro formale fondato sulla teoria dell'informazione parziale, e l'applicazione di questo quadro a problemi filosofici specifici. Conviene ora raccogliere ciò che è stato stabilito, distinguendo con attenzione le tesi centrali dai risultati accessori e dai limiti riconosciuti.

La tesi centrale — la tesi della ridondanza costitutiva propriamente detta — è che il contenuto di uno stato cognitivo è una proprietà relazionale della struttura di coerenza tra le sue rappresentazioni ridondanti, non una proprietà di nessuna rappresentazione individuale. Questa tesi è stata argomentata in tre modi indipendenti: attraverso l'argomentazione dalla necessità funzionale, che mostra che un sistema capace di rappresentazioni stabili e generalizzabili deve avere una struttura di ridondanza costitutiva; attraverso la spiegazione delle dissociazioni neuropsicologiche, che mostra che i dati clinici sono più coerenti con la tesi che con qualsiasi alternativa che mantenga il principio di unicità del portatore; e attraverso l'argomentazione dalla convergenza descrittiva, che mostra che la caratterizzazione formale attraverso la firma informazionale converge con descrizioni sviluppate da tradizioni filosofiche indipendenti.

Il quadro formale fondato sulla teoria dell'informazione parziale ha permesso di precisare la tesi in modo misurabile: la firma informazionale $\Sigma(\Xi, \Theta) = (\mathcal{R}(\Xi, \Theta), \kappa(\Xi, \Theta), \sigma(\Xi, \Theta))$ è l'oggetto matematico che porta il contenuto, e l'identità del contenuto è l'isomorfismo delle firme come strutture reticolari pesate. Questo ha permesso di definire in modo rigoroso la distanza tra contenuti attraverso la metrica di Gromov-Hausdorff, il grado di oggettività del contenuto attraverso la dispersione delle firme nella famiglia di sistemi, la misura dell'errore rappresentazionale come distanza dalla firma di riferimento, e la continuità concettuale come continuità della traiettoria nello spazio delle firme.

Tra i risultati accessori ma significativi vi sono: la tassonomia degli errori rappresentazionali (de-coerenza, soppressione della ridondanza, distorsione sinergica, contaminazione) che offre una classificazione più fine degli errori rispetto alle teorie standard; la caratterizzazione formale del cambiamento concettuale come trasformazione della traiettoria nello spazio delle firme, con la misura di incoerenza ibrida $\mathcal{H}$ per i periodi di transizione; la distinzione formale tra contenuto genuino e contenuto simulato attraverso la distanza di ancoraggio causale Δ_{causal} ; e la misura del disallineamento dei sistemi artificiali come distanza della firma Λ per i concetti normativamente rilevanti.

IX.2 — Confronto con le teorie del contenuto tradizionali: ciò che viene conservato e ciò che viene abbandonato

La tesi della ridondanza costitutiva non è una confutazione globale delle teorie del contenuto esistenti. Ciascuna di esse cattura aspetti reali del fenomeno del contenuto che il quadro della ridondanza costitutiva deve integrare, non eliminare.

La teleosemantica di Millikan cattura il fatto che il contenuto non è arbitrario ma è vincolato dalla storia evolutiva e di sviluppo del sistema cognitivo: la firma informazionale non è una struttura che il sistema può costruire liberamente, ma è modellata dalle pressioni selettive che hanno conformato l'architettura del sistema. Il quadro della ridondanza costitutiva integra questa intuizione riconoscendo che la struttura della firma riflette la storia delle interazioni tra il sistema e l'ambiente — incluse le pressioni selettive — anche se non richiede di fondare la normatività su quella storia.

La teoria informazionale di Dretske cattura il fatto che il contenuto è ancorato causalmente al mondo attraverso relazioni di dipendenza nomologica: la firma informazionale non misura soltanto la struttura interna delle relazioni tra rappresentazioni, ma la struttura delle relazioni tra quelle rappresentazioni e le proprietà del mondo. Il quadro della ridondanza costitutiva integra questa intuizione nella definizione stessa della firma, che è definita rispetto all'obiettivo Θ nel mondo.

La semantica predittiva di Friston cattura il fatto che le rappresentazioni cognitive non sono passive ma attive: il sistema non si limita a registrare le proprietà del mondo ma le predice, e il contenuto è il prodotto dell'inferenza bayesiana sull'input. Il quadro della ridondanza costitutiva integra questa intuizione nel modello dinamico degli attrattori, che descrive la firma come lo stato verso cui il sistema converge attraverso il processo di minimizzazione dell'energia libera.

Ciò che viene abbandonato è la presupposizione di unicità del portatore: l'assunzione che vi sia una singola struttura — un simbolo, una

distribuzione a posteriori, una funzione propria — che costituisce il portatore del contenuto. Questa presupposizione è incompatibile con i dati neurobiologici sulla distribuzione delle rappresentazioni, e la sua eliminazione permette di formulare domande che le teorie basate su di essa non potevano nemmeno porre.

IX.3 — Implicazioni per la filosofia della mente: il naturalismo relazionale

Il contributo più generale della tesi della ridondanza costitutiva al dibattito filosofico più ampio è la proposta di una posizione che si potrebbe chiamare naturalismo relazionale sul contenuto. Questa posizione si colloca in un territorio intermedio tra le tradizionali dicotomie che hanno strutturato il dibattito:

Interno/esterno. Il dibattito tra internalismo e externalismo sul contenuto si è concentrato sulla domanda se il contenuto di uno stato mentale sia determinato dalla struttura interna del sistema o dalle sue relazioni causali con l'ambiente esterno. La firma informazionale dissolve questa dicotomia: è interna perché è una proprietà della struttura del sistema cognitivo, ed è esterna perché è definita rispetto all'oggetto nel mondo. Non vi è un termine privilegiato della relazione — né il sistema né il mondo — ma la struttura della relazione stessa, che è sia interna sia esterna.

Formalismo/naturalismo. Il dibattito tra teorie formali del contenuto (che lo caratterizzano in termini di strutture computazionali astratte) e teorie naturalistiche (che lo radicano in processi fisici e biologici) è attraversato dalla stessa tensione. La firma informazionale è un oggetto formale — una struttura reticolare con misure ben definite — che è al tempo stesso radicata nel fisico attraverso la storia causale dell'interazione con il mondo.

Normatività/naturalismo. Il dibattito tra teorie normativiste (che fondano il contenuto su norme non riducibili ai fatti naturali) e teorie naturaliste (che spiegano la normatività in termini di fatti naturali) riceve dalla firma informazionale una risposta che non privilegia nessuno dei due poli: la norma emerge dalla struttura naturale della firma di riferimento, ma non si riduce a essa in modo brutale — è una proprietà funzionale emergente dall'organizzazione della struttura naturale.

Questa posizione di naturalismo relazionale ha affinità con diverse tradizioni filosofiche. Si avvicina al pragmatismo di Dewey e Peirce nella sua enfasi sulla struttura delle relazioni tra sistema e ambiente piuttosto che sulle proprietà intrinseche degli stati. Si avvicina all'approccio non-rappresentazionalista di Ryle e Wittgenstein nella sua critica della localizzazione del contenuto in strutture interne discrete. Si avvicina all'enattivismo di Maturana, Varela e Thompson nella sua enfasi sulla

costituzione del contenuto attraverso il pattern delle interazioni tra sistema e ambiente.

IX.4 — La fenomenologia come conversa metodologica: lo schema corporeo e la firma informazionale

Una delle convergenze più suggestive che emergono dalla costruzione del quadro della ridondanza costitutiva è quella con la tradizione fenomenologica, in particolare con la descrizione husserliana dell'oggetto intenzionale e con la nozione merleau-pontiana dello schema corporeo.

Nel pensiero di Husserl, l'oggetto intenzionale non è l'oggetto reale nel mondo (l'oggetto fisico) né la rappresentazione interna dell'oggetto (l'immagine mentale), ma il correlato intenzionale dell'atto cognitivo — la struttura che emerge dalla correlazione tra l'atto e l'oggetto. L'oggetto intenzionale è dotato di un «orizzonte» di determinazioni potenziali: è dato secondo certe proprietà nella prospettiva attuale, ma è costituito come struttura che supera la prospettiva attuale attraverso l'anticipazione di prospettive ulteriori. L'identità dell'oggetto intenzionale attraverso prospettive diverse è data dalla struttura di coerenza delle anticipazioni — dal fatto che le diverse prospettive sono mutuamente compatibili e si integrano in un'esperienza unitaria dell'oggetto stesso.

Questa struttura è formalmente analoga alla firma informazionale: l'oggetto intenzionale husserliano è la struttura di coerenza tra le diverse prospettive dell'oggetto, esattamente come la firma informazionale è la struttura di coerenza tra le diverse rappresentazioni ridondanti. Le «prospettive» husserliane corrispondono alle componenti della firma; l'«orizzonte» delle determinazioni potenziali corrisponde alla struttura sinergica della firma; la «coerenza» delle prospettive corrisponde al coefficiente di coerenza κ . La nozione husserliana di «sintesi passiva» — il processo attraverso cui le diverse prospettive si integrano automaticamente in un oggetto percepito come unitario, senza intervento riflessivo consapevole — corrisponde al processo dinamico di convergenza verso l'attrattore della firma descritto nel Capitolo IV.

La nozione merleau-pontiana di schema corporeo approfondisce questa convergenza in una direzione specifica: quella delle componenti motorie e propriocettive del contenuto. Per Merleau-Ponty, il corpo non è semplicemente l'oggetto fisico che occupa lo spazio, ma il «soggetto motore» — il sistema che organizza lo spazio percepito in termini di possibilità motorie. Lo schema corporeo è la struttura che integra le informazioni propriocettive, cinestesiche, tattili e visive in una rappresentazione del corpo che è al tempo stesso il soggetto della percezione e l'oggetto di una percezione di sé. L'identità dello schema

corporeo non risiede in nessuna delle componenti individualmente, ma nella struttura di coerenza che le lega.

Questa struttura corrisponde alla componente motoria-procedurale della firma informazionale, che nel Capitolo II è stata identificata come la terza varietà di ridondanza (epistemico-prospettica) e che nel Capitolo VI ha ricevuto la caratterizzazione formale della normatività implicita. La convergenza tra la descrizione fenomenologica dello schema corporeo e la caratterizzazione formale della componente procedurale della firma è un segnale di coerenza metodologica: due approcci completamente diversi — l'analisi dell'esperienza vissuta e la teoria dell'informazione — convergono verso la medesima struttura di ridondanza tra componenti motorie, propriocettive e visive come base del contenuto della percezione corporea.

IX.5 — Il significato filosofico della tesi: oltre la dicotomia tra interno ed esterno

Il contributo più generale di questa monografia — al di là dei risultati tecnici e delle tesi specifiche che si sono sviluppate — è probabilmente il seguente: avere mostrato che il problema del contenuto mentale non ha bisogno di essere risolto scegliendo tra interno ed esterno, tra naturalismo e normatività, tra struttura formale e ancoraggio al mondo. Le dicotomie che hanno dominato il dibattito — tra teorie che fondano il contenuto sulla struttura interna del sistema e teorie che lo fondano sulle relazioni causali esterne — sono conseguenze della presupposizione di unicità, dell'idea che vi sia una singola relazione o struttura che porta il contenuto. Una volta abbandonata quella presupposizione, le dicotomie perdono la loro presa.

La firma informazionale è interna perché è una proprietà della struttura del sistema cognitivo; è esterna perché è definita rispetto all'oggetto nel mondo e riflette la storia causale delle interazioni. È formale perché ha la struttura di un reticolo misurabile con strumenti matematici precisi; è causalmente ancorata perché la sua struttura riflette la storia delle interazioni del sistema con l'ambiente. È normativa perché permette di misurare la distanza dalla firma di riferimento e di valutare l'adeguatezza delle rappresentazioni; è naturalistica perché la firma di riferimento emerge da processi di apprendimento governati da principi informativi senza appello a norme trascendenti.

La tesi della ridondanza costitutiva non è una sintesi hegeliana delle teorie precedenti — una loro riconciliazione dialettica in una prospettiva superiore. È invece una dissoluzione delle dicotomie attraverso un cambiamento di livello di analisi: il passaggio dal livello delle rappresentazioni singole — dove le dicotomie emergono con forza perché si deve scegliere a quale polo della dicotomia assegnare il portatore del contenuto — al livello delle strutture di relazione tra

rappresentazioni multiple, dove le dicotomie perdono la propria presa perché la struttura di relazione è intrinsecamente sia interna sia esterna, sia formale sia causale, sia normativa sia naturale.

Questo cambiamento di livello ha un costo: richiede un apparato formale più ricco di quello impiegato dalle teorie precedenti — la teoria matematica delle dipendenze multivariate, la geometria dell'informazione, la topologia persistente — che è estraneo alla tradizione della filosofia del linguaggio e della filosofia analitica della mente. Ma ha un guadagno che giustifica il costo: permette di formulare domande che le teorie precedenti non potevano nemmeno porre — domande sulla struttura della ridondanza, sulla sinergia, sulla distanza tra firme, sulla continuità delle traiettorie nello spazio dei contenuti — e di rispondervi in modo empiricamente testabile.

Il contenuto mentale, nella prospettiva della ridondanza costitutiva, non è un mistero filosofico che la scienza non può toccare né un fatto empirico che la filosofia non deve problematizzare. È una struttura che attraversa entrambi i domini, che si misura con gli strumenti della teoria dell'informazione e si articola con gli strumenti della filosofia analitica, e che abita — come tutte le strutture che contano davvero — molti luoghi simultaneamente.

Capitolo X — La linguistica cognitiva come convergenza indipendente: frame, prototipi e semantica radicale

X.1 — Tre tradizioni, un problema comune

Esiste, nella linguistica cognitiva dei decenni compresi tra la metà degli anni Settanta e la fine del Novecento, una convergenza che è rimasta largamente inosservata nella filosofia del contenuto mentale. Tre programmi di ricerca sviluppati con metodologie diverse, da ricercatori che lavoravano in contesti accademici spesso privi di connessioni dirette tra loro, sono approdati a risultati che, visti retrospettivamente, costituiscono una conferma indipendente e convergente della tesi della ridondanza costitutiva del contenuto. Questi programmi sono la semantica dei frame di Charles Fillmore, la teoria dei prototipi di Eleanor Rosch, e la semantica cognitiva radicale del programma di George Lakoff e Mark Johnson. La convergenza non è superficiale né è riducibile a mere somiglianze terminologiche: è una convergenza nel merito, che tocca la struttura del contenuto lessicale, le condizioni di categorizzazione, e la natura delle relazioni semantiche tra concetti.

La tesi comune che emerge da questi tre programmi — espressa con terminologie e strumenti diversi — è che il significato di una parola o di

un concetto non è determinato da un insieme di condizioni necessarie e sufficienti, né da un singolo prototipo isolato, né da una relazione causale a due posti tra la parola e la proprietà nel mondo. Il significato è determinato da una struttura relazionale complessa che coinvolge simultaneamente molte dimensioni semantiche, molti tipi di conoscenza di sfondo, molti schemi di classificazione alternativa, molte prospettive esperienziali. In altri termini, il significato ha una struttura distribuita e multipla che è riconoscibile come una versione — sviluppata in chiave linguistica piuttosto che informatica o neuroscientifica — della struttura di ridondanza costitutiva qui difesa.

La connessione non è soltanto analogica. Come si mostrerà nel corso di questo capitolo, i costrutti tecnici della semantica dei frame, della struttura prototipica e degli schemi di immagine si lasciano tradurre con relativa precisione nei termini della firma informazionale: i frame di Fillmore corrispondono alle strutture di ridondanza che definiscono gli atomi informativi di ordine superiore; la struttura prototipica di Rosch corrisponde alla distribuzione dell'informazione unica e ridondante intorno a exemplar centrali; gli schemi di immagine di Lakoff corrispondono alle componenti sinergiche emergenti dai pattern di coattivazione sensomotoria. Questa traduzione non è una semplice mappatura terminologica: è una riduzione teorica parziale che mostra come i fenomeni semantici osservati dalla linguistica cognitiva possano essere spiegati nell'ambito del quadro formale sviluppato in questa monografia.

X.2 — La semantica dei frame di Fillmore: il contenuto come struttura contestuale obbligata

X.2,1 — Dalla valenza verbale ai frame semantici

Charles Fillmore è arrivato alla nozione di frame semantico attraverso un percorso che inizia dalla linguistica strutturale della valenza verbale e approda a una teoria del significato radicalmente contestuale. Il punto di partenza è il concetto di valenza verbale: ogni verbo impone ai propri argomenti una struttura di ruoli tematici — agente, paziente, tema, meta, fonte, strumento — che determina come i partecipanti all'evento verbale si relazionano tra loro e all'evento stesso. Nei lavori degli anni Sessanta e Settanta, Fillmore ha sviluppato la nozione di caso profondo come categoria universale che cattura queste relazioni tematiche a un livello di astrazione superiore alla superficie grammaticale delle diverse lingue.

La svolta verso la semantica dei frame avviene negli anni Ottanta, quando Fillmore riconosce che i ruoli tematici della valenza verbale sono insufficienti per caratterizzare il significato lessicale in modo adeguato. Il problema è che le parole non solo predicano ruoli tematici ma attivano

intere strutture di conoscenza di sfondo — strutture che determinano le implicazioni, le presupposizioni e le condizioni di uso appropriato della parola, senza essere direttamente assertite dal suo contenuto letterale. Una parola come «lunedì» non è comprensibile senza attivare l'intera struttura della settimana: i giorni che precedono e seguono, la loro distribuzione nella struttura lavorativa/non-lavorativa della cultura d'uso, le loro relazioni con i cicli commerciali e liturgici. Una parola come «ipoteca» non è comprensibile senza attivare il frame delle transazioni finanziarie a lungo termine, con la struttura delle parti coinvolte (mutuante, mutuatario, creditore, garante), degli oggetti rilevanti (il bene immobile, il contratto, la rata), degli eventi pertinenti (la concessione, il pagamento, il default, l'estinzione). Il significato di «ipoteca» è, in questa prospettiva, non il contenuto isolato del lemma ma la posizione del lemma nella struttura del frame delle transazioni finanziarie.

Questa idea ha una conseguenza teorica di grande portata: il significato di una parola non è determinabile in isolamento ma solo in relazione al frame che quella parola evoca. Due parole che appartengono allo stesso frame hanno significati connessi da relazioni strutturali che non sono catturabili dall'analisi dei singoli lessemi: «compratore» e «venditore» non sono due concetti indipendenti che si trovano ad avere una relazione funzionale, ma sono le due facce di una relazione commerciale unificata che il frame della transazione commerciale codifica. Il significato di «compratore» presuppone necessariamente l'esistenza di un venditore, di un oggetto compravenduto, di un pagamento, di un contesto di accordo: questi elementi non sono implicazioni pragmatiche che si aggiungono al significato letterale, ma sono parte costitutiva del significato stesso, perché è solo all'interno del frame che il concetto di compratore ha senso.

X.2,2 — I frame come strutture di ridondanza semantica

La connessione con la ridondanza costitutiva si manifesta nel modo seguente. Un frame, nella terminologia di Fillmore, è una struttura cognitiva che organizza la conoscenza relativa a un dominio di esperienze. Il frame della «vendita» organizza i concetti di venditore, acquirente, merce, prezzo, pagamento, transazione, accordo in una struttura in cui ciascun elemento è definito dalle sue relazioni con gli altri. Non vi è un elemento del frame che sia semanticamente primitivo e da cui gli altri derivino per definizione: tutti gli elementi si definiscono reciprocamente, formando una struttura di co-dipendenze semantiche. Questa struttura è precisamente una struttura di ridondanza costitutiva nel senso della firma informazionale: ciascun elemento del frame porta informazione che è parzialmente ridondante rispetto agli altri elementi (tutti si riferiscono allo stesso evento di vendita), parzialmente unica (ciascuno cattura un aspetto specifico dell'evento: il venditore non coincide semanticamente con l'acquirente), e parzialmente sinergica (il significato della transazione come evento unificato emerge solo dalla

struttura delle relazioni tra tutti gli elementi, non da nessuno di essi individualmente).

Per rendere questa connessione più precisa, consideriamo la decomposizione PID applicata al frame della vendita. Siano X_1, X_2, X_3, X_4 le sorgenti informatiche corrispondenti agli elementi del frame (venditore, acquirente, merce, prezzo), e sia Y il contenuto «vendita» come evento unificato. La struttura delle relazioni informative tra X_1 - X_4 rispetto a Y determina la firma informazionale del frame:

$$\Sigma(X_1, X_2, X_3, X_4 ; Y) = \{R(\text{frame}), U_{\{X_i\}}(\text{frame}), S_2(\text{coppie}), S_3(\text{triple}), S_4(\text{quadrupla})\}$$

La ridondanza $R(\text{frame})$ è la componente dell'informazione sulla vendita che è condivisa da tutti gli elementi del frame: è l'informazione che indica che si tratta di un evento di scambio commerciale, informazione che è presente in ciascuno degli elementi del frame e che non dipende dalla struttura specifica delle loro relazioni. Le informazioni uniche $U_{\{X_i\}}$ sono gli aspetti specifici di ciascun elemento che non sono presenti negli altri: l'informazione sull'identità del venditore, sull'identità dell'acquirente, sulle caratteristiche della merce, sull'ammontare del prezzo. Le sinergie di ordine 2 sono le informazioni che emergono solo dall'analisi congiunta di coppie di elementi: la relazione di scambio emerge solo dall'analisi congiunta di venditore e acquirente, il rapporto qualità-prezzo emerge solo dall'analisi congiunta di merce e prezzo. La sinergia di ordine 4 è il contenuto dell'intera transazione come evento unificato, che non emerge dalla somma delle componenti né da nessuna combinazione di tre elementi.

Fillmore ha costruito una banca dati lessicale — FrameNet, sviluppata a partire dagli anni Novanta con il supporto dell'Università della California a Berkeley — che organizza il lessico dell'inglese (e successivamente di molte altre lingue) in termini di frame semantici e degli elementi di frame che le parole possono evocare. FrameNet è, in questo senso, una mappatura empirica della struttura dei frame semantici del lessico inglese, e può essere letta come una mappatura parziale della struttura delle firme informative associate ai domini di conoscenza organizzati dalla lingua.

X.2,3 — La teoria della prospettiva semantica e la variabilità della firma

Un contributo ulteriore della semantica dei frame che è direttamente rilevante per il quadro della ridondanza costitutiva è la teoria della prospettiva semantica. Fillmore ha osservato che molti frame semantici possono essere «profilati» da prospettive diverse: gli elementi del frame possono essere messi a fuoco selettivamente in modo che la stessa struttura di conoscenza sia espressa linguisticamente in modo diverso. Il caso paradigmatico è quello dei verbi di moto: un verbo come

«comprare» profila la prospettiva dell'acquirente nel frame della transazione commerciale (l'acquirente è in primo piano, il venditore è sullo sfondo), mentre un verbo come «vendere» profila la prospettiva del venditore (il venditore è in primo piano, l'acquirente è sullo sfondo), e un verbo come «costare» profila la relazione tra merce e prezzo senza focalizzare né venditore né acquirente. Tutti e tre i verbi evocano il medesimo frame — la transazione commerciale — ma lo profilano da prospettive diverse.

Questa nozione di profilazione corrisponde, nella terminologia della firma informazionale, alla selezione di diverse componenti della firma come figura rispetto allo sfondo delle altre. Quando si usa «comprare», le componenti della firma associate all'acquirente sono in primo piano: la loro informazione unica e le loro relazioni sinergiche con gli altri elementi del frame sono semanticamente più salienti. Quando si usa «vendere», le componenti associate al venditore sono in primo piano. La stessa firma informazionale del frame può essere profilata in modi diversi, producendo differenze semantiche sistematiche che non corrispondono a differenze nel contenuto del frame ma a differenze nell'organizzazione della sua struttura saliente.

Questo risultato ha un'implicazione importante: la struttura della firma informazionale del contenuto non è solo un oggetto semantico ma ha anche una struttura di prominenza differenziale — una struttura di figura e sfondo nel senso della psicologia della Gestalt — che corrisponde alle diverse possibilità di profilazione semantica. Una teoria del contenuto basata sulla firma deve dunque includere non solo la struttura delle relazioni informatiche tra le componenti, ma anche la struttura della prominenza differenziale che determina quali componenti sono profilate come figura in un dato contesto d'uso.

X.3 — La teoria dei prototipi di Rosch: la categorizzazione come struttura di somiglianza familiare

X.3,1 — Il programma roscheano e il crollo delle condizioni necessarie e sufficienti

Eleanor Rosch ha sviluppato la teoria dei prototipi nel corso degli anni Settanta come risposta empirica alla teoria classica della categorizzazione. La teoria classica — articolata con varie sfumature da Aristotele fino ai logici analitici del Novecento — sostiene che ogni categoria è definita da un insieme di condizioni necessarie e sufficienti: un oggetto appartiene alla categoria C se e solo se soddisfa tutte le condizioni necessarie per C, e queste condizioni sono congiuntamente sufficienti per l'appartenenza. In questa visione, l'appartenenza a una categoria è binaria e tutte le istanze della categoria sono ugualmente istanze: un pettirosso è un

uccello esattamente quanto uno struzzo, poiché entrambi soddisfano le condizioni necessarie e sufficienti per l'appartenenza alla categoria degli uccelli.

L'evidenza sperimentale raccolta da Rosch in una serie di esperimenti sistematici condotti su molte categorie semantiche di diversi livelli di generalità ha mostrato che questo quadro è empiricamente sbagliato in modo piuttosto radicale. Gli esperimenti di Rosch hanno rivelato due fatti che la teoria classica non può accomodare. Il primo fatto è quello della tipicità gradata: all'interno di una categoria, i membri non sono ugualmente tipici. Un pettirosso è giudicato dai parlanti una istanza più tipica di «uccello» di quanto non lo siano uno struzzo, un pinguino o un pipistrello. Il pettirosso è un prototipo — un'istanza centrale della categoria — mentre lo struzzo è un'istanza periferica. Questo gradiente di tipicità non è riducibile alla frequenza di esposizione né a preferenze culturali arbitrarie: è sistematico e stabile tra popolazioni diverse.

Il secondo fatto è quello della struttura di somiglianza familiare: i membri di una categoria non condividono necessariamente una proprietà comune a tutti — quella che sarebbe la condizione necessaria e sufficiente dell'appartenenza — ma condividono overlapping clusters di proprietà nel senso del «familienähnlichkeit» di Wittgenstein. Il pettirosso e il passero condividono la proprietà di volare, cantare e avere una certa dimensione corporea; il pinguino condivide con il passero la dimensione corporea ma non il volo né il canto; lo struzzo condivide con il pinguino la non-capacità di volare ma non la dimensione corporea né il canto. Non vi è una proprietà unica che tutti e quattro condividano, eppure tutti appartengono alla categoria degli uccelli.

Rosch ha spiegato questa struttura introducendo la nozione di prototipo come struttura cognitiva che organizza la conoscenza categoriale: il prototipo è la rappresentazione dell'istanza più tipica della categoria, e la tipicità di un'istanza specifica è determinata dalla sua similarità con il prototipo. La categorizzazione non è un processo di verifica di condizioni necessarie e sufficienti ma è un processo di calcolo della somiglianza con il prototipo.

X.3,2 — Il prototipo come struttura informazionale centrale

La teoria dei prototipi di Rosch, nella sua interpretazione più forte, ha incontrato critiche serie che hanno portato a raffinamenti importanti. La critica più rilevante per la nostra discussione è quella di Sharon Armstrong, Lila Gleitman e Henry Gleitman, i quali hanno mostrato che i giudizi di tipicità si osservano anche per categorie geometriche definite da condizioni necessarie e sufficienti: i soggetti giudicano che un numero pari come 4 è un numero pari più tipico di un numero pari come 18, anche se entrambi soddisfano identicamente la condizione necessaria e sufficiente dell'essere divisibili per 2. Questo risultato suggerisce che i

giudizi di tipicità riflettono non la struttura categoriale in senso stretto ma qualcosa come la «goodness of example» — la qualità di esemplare, che dipende da fattori come la salienza cognitiva, la semplicità e la frequenza di esposizione — e che questa struttura di tipicità è separabile dalla struttura categoriale vera e propria.

Il raffinamento della teoria dei prototipi che meglio sopravvive a questa critica è quello proposto da Lawrence Barsalou, il quale ha introdotto la nozione di concetto situato in sostituzione del concetto come rappresentazione stabile. Nella teoria di Barsalou, i concetti non sono strutture cognitive fisse ma vengono costruiti dinamicamente in risposta al contesto d'uso: la rappresentazione attivata per il concetto UCCELLO varia sistematicamente a seconda del contesto in cui il concetto è utilizzato — un uccello che si mangia (pollo arrosto), un uccello che si osserva (pettirosso che canta), un uccello che si teme (aquila in picchiata), un uccello che si alleva (canarino in gabbia). La struttura prototipica non è, in questa prospettiva, una struttura cognitiva fissa bensì un attrattore dinamico che si manifesta diversamente in contesti diversi.

Questa revisione barsalouviana della teoria dei prototipi è particolarmente illuminante per il quadro della firma informazionale. Il concetto come costruzione situata è precisamente un sistema in cui la firma informazionale varia con il contesto: componenti diverse della firma vengono profilate in modo diverso in contesti d'uso diversi, e la struttura di ridondanza, unicità e sinergia tra le componenti cambia sistematicamente con il contesto. Un uccello-da-mangiare ha una firma informazionale centrata sulle componenti tattile-gustative e sulle relazioni con il frame culinario; un uccello-da-osservare ha una firma centrata sulle componenti visivo-uditivo e sulle relazioni con il frame naturalistico. La stessa struttura di conoscenza categoriale produce firme diverse in contesti diversi, e la struttura del concetto situato è la struttura di questo spazio di firme possibili più la distribuzione probabilistica sulle condizioni di profilazione contestuale.

X.3,3 — La struttura radiale di Lakoff e la famiglia di estensioni

George Lakoff ha sviluppato il concetto di modello cognitivo idealizzato (MCI, Idealized Cognitive Model) come generalizzazione della teoria dei prototipi, e ha introdotto la nozione di struttura radiale per descrivere le categorie con struttura prototipica complessa e molte estensioni periferiche. Una struttura radiale ha un centro — il prototipo o il caso centrale — e molte estensioni che si dipartono dal centro attraverso diversi tipi di trasformazione: metonimia, metafora, specializzazione, generalizzazione, o ristrutturazione di prospettiva. La categoria MADRE, analizzata da Lakoff come esempio paradigmatico, ha un caso centrale (la madre biologica che alleva il figlio) e molte estensioni: la madre adottiva (non biologica ma allevatrice), la madre surrogata (biologica ma non allevatrice), la madre biologica non allevatrice (come nei casi di adozione

o abbandono), la madre foster, la madre genetica (donatrice di ovulo). Nessuna di queste estensioni è la «vera» madre nel senso delle condizioni necessarie e sufficienti: ciascuna cattura alcuni ma non tutti gli aspetti del caso centrale.

La struttura radiale di Lakoff è, nella terminologia della firma informazionale, una struttura in cui la firma del caso centrale ha una configurazione specifica (alta ridondanza tra le componenti biologica, allevatrice, familiare, emozionale) e le firme delle estensioni periferiche si ottengono dalla firma centrale per diminuzione sistematica di specifiche componenti di ridondanza: la madre adottiva ha bassa ridondanza sulla componente biologica ma alta ridondanza sulle componenti allevatrice e familiare; la madre surrogata ha alta ridondanza sulla componente biologica ma bassa sulla componente allevatrice. La struttura radiale è la struttura di questo spazio di firme derivate dalla firma centrale per trasformazioni sistematiche delle componenti di ridondanza.

Questa caratterizzazione formale della struttura radiale ha un vantaggio esplicativo: spiega perché le estensioni periferiche delle categorie radiali sono sistematiche piuttosto che arbitrarie. Non qualsiasi manipolazione della firma del caso centrale produce un'estensione semanticamente accettabile: solo le manipolazioni che preservano un grado sufficiente di sovrapposizione informativa con la firma centrale producono estensioni che i parlanti riconoscono come varianti della stessa categoria. La soglia di accettabilità è determinata dalla distanza di Gromov-Hausdorff tra la firma dell'estensione proposta e la firma del caso centrale:

$$d_{GH}(\Sigma(\text{madre_X}), \Sigma(\text{madre_centrale})) < \delta$$

dove δ è la soglia di riconoscimento categoriale, che varia per tipologie diverse di categoria e per contesti culturali diversi. Le categorie con struttura radiale densa (molte estensioni accettabili) sono quelle con firme centrali ad alta entropia relativa — firme che contengono molti atomi informativi di cui ciascuno può variare indipendentemente producendo firme periferiche ancora riconoscibili. Le categorie con struttura radiale sparsa (poche estensioni accettabili) sono quelle con firme centrali a bassa entropia relativa — firme in cui la coerenza tra gli atomi è così alta che la rimozione di anche uno solo di essi produce una firma troppo distante per essere riconosciuta come estensione.

X.3,4 — I livelli di base della categorizzazione e la struttura ottimale della firma

Uno dei contributi più importanti della teoria dei prototipi di Rosch — e uno di quelli con le implicazioni più dirette per la teoria della firma informazionale — è la teoria dei livelli di base della categorizzazione. Rosch, Berlin e Kay hanno mostrato che i sistemi di categorizzazione naturale delle lingue umane non si distribuiscono in modo uniforme su tutti i livelli di generalità possibili, ma tendono a privilegiare un livello

intermedio — il livello di base — che ha proprietà psicologiche e comunicative ottimali.

Il livello di base è il livello in cui le categorie sono massimamente informative rispetto alla dimensione cognitiva: il livello in cui le istanze dello stesso livello condividono il massimo di attributi comuni e le istanze di livelli diversi condividono il minimo di attributi comuni. «Cane» è un livello di base; «animale» è un livello superordinato più astratto; «labrador» è un livello subordinato più specifico. «Cane» è cognitivamente privilegiato perché i cani condividono molti attributi tra loro (hanno quattro zampe, fanno «bau», hanno pelo, si possono addestrare) e condividono pochi attributi con i gatti, i cavalli e i pesci. Al livello superordinato «animale», le istanze condividono pochi attributi (le caratteristiche fondamentali della vita animale), il che riduce l'informatività categoriale. Al livello subordinato «labrador», si guadagna specificità ma si perde generalizzabilità perché molti attributi del labrador non si generalizzano agli altri cani.

Nella teoria della firma informazionale, il livello di base corrisponde al livello in cui la firma informazionale ha la configurazione ottimale: massima ridondanza tra le componenti (gli esemplari del livello di base condividono molta informazione), massima unicità delle componenti rispetto ai livelli adiacenti (la firma del livello di base è distinguibile dalle firme dei livelli superordinato e subordinato), e sufficiente sinergia per catturare le relazioni strutturali tra gli esemplari che vanno oltre la semplice condivisione di proprietà. Formalmente, il livello di base L^* massimizza una funzione della forma:

$$L^* = \operatorname{argmax}_L [\kappa(L) \cdot (1 + \sigma(L)) / d_{GH}(\Sigma(L), \Sigma(L_{sup})) \cdot d_{GH}(\Sigma(L), \Sigma(L_{sub}))]$$

dove $\kappa(L)$ è il numero di atomi ridondanti nella firma al livello L , $\sigma(L)$ è la media delle componenti di sinergia, $\Sigma(L_{sup})$ e $\Sigma(L_{sub})$ sono le firme dei livelli superordinato e subordinato, e i due fattori al denominatore misurano la distanza della firma del livello L dalle firme dei livelli adiacenti. Il livello di base è quello in cui la firma ha alta coerenza interna (κ alto), alta ricchezza (σ alto), e alta distinguibilità dai livelli adiacenti (fattori al denominatore grandi).

Questa derivazione formale del livello di base dalla struttura della firma informazionale è un risultato non banale: spiega per quale ragione il livello di base è cognitivamente privilegiato non in termini di utilità comunicativa o frequenza di uso (spiegazioni che sono corrette ma superficiali), ma in termini della struttura delle dipendenze informatiche tra le rappresentazioni delle categorie a livelli diversi. Il livello di base è privilegiato perché corrisponde al livello in cui la struttura informazionale delle rappresentazioni è massimamente integrata: il livello in cui la firma ha la configurazione di massima coerenza, e in cui quindi il

contenuto è più robusto, più generalizzabile e più resistente a perturbazioni contestuali.

X.4 — La semantica radicale di Lakoff e Johnson: schemi di immagine e struttura sensomotoria del contenuto

X.4,1 — Il programma della linguistica cognitiva di Lakoff

George Lakoff, insieme a Mark Johnson, ha elaborato nel corso degli anni Ottanta e Novanta un programma di ricerca che ha il suo punto di partenza nel libro *Metaphors We Live By* del 1980 e si sviluppa nella grande sintesi teorica di *Women, Fire, and Dangerous Things* del 1987 e *Philosophy in the Flesh* del 1999. Il progetto di Lakoff è radicalmente anti-oggettivista: contro la tradizione logicista che considera il significato come una relazione tra simboli e condizioni di verità indipendenti dalla mente, Lakoff afferma che il significato è incorporato — è strutturato dall'esperienza senso motoria del corpo che interagisce con il suo ambiente fisico e sociale.

Il nucleo della posizione lakoffiana è la tesi degli schemi di immagine come strutture cognitive fondamentali che organizzano il contenuto concettuale. Gli schemi di immagine non sono immagini nel senso di rappresentazioni visive ricche di dettaglio: sono strutture schematiche astratte derivate da pattern ricorrenti di esperienza corporea. Lo schema CONTENITORE — formato dall'esperienza ricorrente di essere dentro o fuori da oggetti fisici, di mettere oggetti in contenitori, di muoversi dentro e fuori spazi delimitati — organizza una vasta gamma di concetti astratti: categorie mentali (appartenere o non appartenere a una categoria è un'esperienza di «dentro» o «fuori»), stati emotivi («essere nella depressione», «uscire dalla crisi»), relazioni sociali («essere nel partito», «escludere qualcuno dal gruppo»). Lo schema PERCORSO — formato dall'esperienza ricorrente di muoversi da un punto di origine verso una meta attraverso un tragitto — organizza i concetti di causa ed effetto (la causa è l'origine, l'effetto è la meta), di intenzione (l'intenzione è la meta, l'azione è il percorso), di cambiamento di stato (lo stato iniziale è l'origine, lo stato finale è la meta).

X.4,2 — Schemi di immagine come strutture di sinergia sensomotoria

La connessione con la ridondanza costitutiva si manifesta con forza nella struttura degli schemi di immagine come strutture di coattivazione sensomotoria. Uno schema di immagine come CONTENITORE non è un'unica rappresentazione: è una struttura di ridondanza tra molte rappresentazioni sensorimotorie diverse che coattivano in modo

sistematico e che producono, attraverso la loro coattivazione, un contenuto che non appartiene a nessuna di esse individualmente.

Consideriamo in dettaglio la struttura del schema CONTENITORE. La coattivazione sensomotiva che produce questo schema include: la rappresentazione visiva di una frontiera che separa uno spazio interno da uno spazio esterno (componente L_1), la rappresentazione tattile della resistenza fisica di una parete (L_2), la rappresentazione propriocettiva del confine del proprio corpo come delimitazione tra il sé e l'ambiente (L_3), la rappresentazione cinestetica del movimento di attraversamento di un confine (L_4), la rappresentazione della sensazione di protezione o confinamento quando si è all'interno (L_5), e la rappresentazione mnemonica degli eventi ricorrenti di mettere e togliere oggetti da contenitori (L_6). La firma informazionale di CONTENITORE come schema cognitivo è la struttura delle dipendenze informatiche tra L_1 - L_6 :

$$\Sigma(\text{CONTENITORE}) = \{R, U_{\{L_1\}}, \dots, U_{\{L_6\}}, S_2(\text{coppie}), S_3(\text{triple}), \dots\}$$

La componente sinergica di CONTENITORE emerge dall'analisi congiunta di L_1 - L_6 : l'idea di un confine fisico che separa un interno da un esterno, con le conseguenti strutture di appartenenza/non-appartenenza, di protezione/esclusione, di entrata/uscita, non è riducibile a nessuna delle componenti sensorimotorie individuali né alla loro somma. È il contenuto del pattern globale di coattivazione, e corrisponde esattamente alla struttura di sinergia di ordine superiore che la PID identifica.

La proiezione metaforica degli schemi di immagine su domini astratti — il meccanismo fondamentale della semantica lakoffiana — è interpretabile come estensione della firma informazionale dello schema sensomotorio al dominio astratto. Quando diciamo che una persona «è nella depressione», stiamo applicando la firma informazionale di CONTENITORE (con le sue componenti di delimitazione, confinamento, protezione/prigionia, entrata/uscita) al dominio degli stati emotivi. La proiezione non è arbitraria: le componenti della firma di CONTENITORE che vengono proiettate sulla depressione sono quelle che catturano gli aspetti dell'esperienza della depressione che hanno struttura analoga all'esperienza dei contenitori fisici (il senso di essere confinati, la difficoltà di uscire, la delimitazione dello spazio emotivo disponibile). Le componenti che non hanno analogia pertinente nel dominio della depressione (la resistenza tattile della parete, il colore visivo del confine) non vengono proiettate.

Questa selettività della proiezione metaforica è spiegabile in termini di preservazione delle strutture sinergiche rilevanti: la metafora proietta le componenti della firma sorgente che hanno relazioni sinergiche strutturalmente analoghe a quelle che le componenti del dominio bersaglio devono catturare. Formalmente, la proiezione metaforica M da

un dominio sorgente S a un dominio bersaglio T è un morfismo parziale della firma:

$$M: \Sigma(S) \rightarrow \Sigma(T)$$

che preserva le relazioni sinergiche strutturalmente pertinenti e scarta le componenti sensorimotorie concrete che non hanno analoghi nel dominio astratto. La qualità di una metafora — la sua coerenza, il suo potere esplicativo, la sua estendibilità a nuovi casi — è determinata dalla struttura del morfismo M : una metafora coerente è una in cui M è un morfismo fedele che preserva molte relazioni sinergiche della firma sorgente, producendo nel dominio bersaglio una struttura di ridondanza analoga a quella del dominio sorgente.

X.4,3 — La semantica radicale contro il mentalismo classico

Il programma di Lakoff e Johnson è esplicitamente critico nei confronti di ciò che chiamano il «mentalismo classico» — la concezione secondo cui il significato è determinato da relazioni tra simboli mentali e il mondo, senza mediazione corporea. La critica lakoffiana al mentalismo classico è convergente con la critica alla presupposizione di unicità sviluppata in questo lavoro, anche se i percorsi argomentativi sono diversi.

Per Lakoff e Johnson, il mentalismo classico sbaglia nel presupporre che il contenuto concettuale sia indipendente dall'esperienza corporea: i concetti non sono simboli astratti che fluttuano o liberi dalle loro radici sensorimotorie, ma sono strutturati dall'esperienza corporea in modo costitutivo. Per la teoria della ridondanza costitutiva, il mentalismo classico sbaglia nel presupporre che il contenuto sia portato da singoli stati localizzati: il contenuto è una struttura relazionale che emerge dalla firma informazionale del sistema. Queste due critiche si rafforzano a vicenda: se il contenuto è incorporato, allora la struttura sensomotoria della firma è parte costitutiva del contenuto, non un mero fattore di modulazione; se il contenuto è una struttura relazionale emergente, allora la struttura incorporata dell'esperienza sensomotoria è esattamente il tipo di struttura relazionale complessa — con le sue componenti di ridondanza, unicità e sinergia tra modalità diverse — che costituisce la firma del contenuto.

X.5 — La semantica dei prototipi e la categorizzazione: convergenza con la firma informazionale

X.5,1 — Il problema della generalizzazione categoriale

Una delle lacune più significative nella trattazione standard della categorizzazione è la difficoltà di spiegare la generalizzazione categoriale — la capacità di riconoscere nuove istanze di una categoria che non assomigliano prototipicamente alle istanze già conosciute. Le teorie basate su condizioni necessarie e sufficienti trattano la generalizzazione come applicazione di un algoritmo di verifica: un nuovo oggetto appartiene alla categoria se soddisfa le condizioni. Le teorie basate su prototipi trattano la generalizzazione come calcolo di similarità: un nuovo oggetto appartiene alla categoria se è sufficientemente simile al prototipo.

Entrambe le teorie trovano difficoltà quando le istanze nuove si discostano sia dalle condizioni classiche sia dai prototipi in modi sistematici ma non previsti. Il caso paradigmatico è quello delle anomalie categoriali sistematiche: oggetti che appartengono chiaramente a una categoria ma che non soddisfano né le condizioni necessarie e sufficienti né il criterio di similarità con il prototipo. Un uccello senza piume, privato delle piume da una malattia, continua a essere classificato come uccello anche se non soddisfa la condizione delle piume e non assomiglia al prototipo piumato. Un cane senza zampe, per una mutilazione, continua a essere classificato come cane anche se non soddisfa la condizione delle zampe e non assomiglia al prototipo quadrupede.

La spiegazione di questi casi attraverso la firma informazionale è più soddisfacente. Un oggetto viene categorizzato come membro di una categoria non perché soddisfi condizioni necessarie e sufficienti, né perché sia simile a un prototipo fisso, ma perché la sua firma informazionale è sufficientemente vicina (nel senso della distanza di Gromov-Hausdorff) alla firma della categoria. La firma della categoria non è la firma di un singolo prototipo ma è la firma del sistema di rappresentazioni che il soggetto ha sviluppato per la categoria nel corso della sua esperienza, e questa firma include componenti di ridondanza che catturano le regolarità statistiche dell'esperienza categoriale, componenti di unicità che catturano gli aspetti idiosincratici degli esemplari incontrati, e componenti sinergiche che catturano le relazioni strutturali tra gli aspetti diversi degli esemplari.

Un uccello senza piume ha una firma che si discosta dalla firma della categoria in alcune componenti (quelle relative alle piume), ma rimane vicina nelle componenti che catturano l'anatomia scheletrica degli uccelli, il loro profilo comportamentale, la loro biologia riproduttiva, le loro relazioni ecologiche. La distanza di Gromov-Hausdorff tra la firma dell'uccello senza piume e la firma della categoria è sufficientemente piccola da permettere la categorizzazione, anche se la firma non coincide con quella del prototipo.

X.5,2 — L'effetto di contesto nelle gerarchie categoriali

Un fenomeno ulteriore che la teoria della firma informazionale spiega più adeguatamente delle teorie classiche è l'effetto di contesto nelle gerarchie categoriali. Diversi esperimenti hanno mostrato che la categoria a cui un oggetto viene assegnato dipende dal contesto categoriale in cui l'oggetto è presentato: lo stesso oggetto può essere classificato come «coltello» in un contesto culinario, come «arma» in un contesto di sicurezza, come «strumento» in un contesto artigianale. Questa variabilità contestuale non è spiegabile né dalla teoria classica delle condizioni necessarie e sufficienti (le condizioni non variano con il contesto) né dalla teoria del prototipo fisso (il prototipo non varia con il contesto).

La spiegazione in termini di firma informazionale è la seguente. L'oggetto ha una firma informazionale ricca che include componenti corrispondenti a molte dimensioni di classificazione possibile: la sua forma fisica (lama affilata, manico), la sua funzione (tagliare), il suo impatto potenziale (possibilità di ferire), il contesto d'uso tipico (cucina, officina). In diversi contesti, diverse componenti della firma sono profilate come figura: nel contesto culinario, le componenti legate alla funzione di taglio alimentare e al contesto d'uso in cucina sono in primo piano; nel contesto di sicurezza, le componenti legate all'impatto potenziale e alla pericolosità sono in primo piano; nel contesto artigianale, le componenti legate alla precisione dello strumento e al contesto artigianale sono in primo piano. La categorizzazione non è determinata da una firma fissa ma dalla firma profilata nel contesto specifico, e la profilazione è determinata dalla struttura di salienza del contesto che amplifica selettivamente alcune componenti della firma.

Formalmente, sia $\Sigma_full(x)$ la firma completa dell'oggetto x e sia C un contesto di categorizzazione. La firma profilata nel contesto C è:

$$\Sigma_C(x) = P_C \circ \Sigma_full(x)$$

dove P_C è un operatore di profilazione che seleziona le componenti della firma rilevanti per il contesto C e aumenta la loro salienza relativa. La categorizzazione avviene confrontando $\Sigma_C(x)$ con le firme delle categorie candidate nel contesto C : x viene assegnato alla categoria K^* che minimizza la distanza:

$$K^* = \operatorname{argmin}_K d_GH(\Sigma_C(x), \Sigma_C(K))$$

dove $\Sigma_C(K)$ è la firma della categoria K profilata nel contesto C . Questo meccanismo spiega la variabilità contestuale della categorizzazione come conseguenza della variabilità della profilazione delle firme, senza fare appello a inconsistenze della struttura categoriale o a ambiguità nel significato delle parole.

X.6 — Convergenze e divergenze tra linguistica cognitiva e ridondanza costitutiva

X.6,1 — Tre convergenze strutturali

Le tre tradizioni di linguistica cognitiva esaminate convergono con la tesi della ridondanza costitutiva in tre punti strutturali che meritano di essere esplicitati.

La prima convergenza riguarda la strutturalità relazionale del contenuto. In tutte e tre le tradizioni — quella dei frame di Fillmore, quella dei prototipi di Rosch, quella degli schemi di immagine di Lakoff — il contenuto semantico di una parola o di un concetto non è determinabile isolatamente ma solo in relazione a una struttura più ampia: la struttura del frame (Fillmore), la struttura della distribuzione degli esemplari intorno al prototipo (Rosch), la struttura della coattivazione sensomotoria (Lakoff). Questa strutturalità relazionale corrisponde alla struttura relazionale della firma informazionale: il contenuto è la firma, non il singolo portatore.

La seconda convergenza riguarda la gradualità del contenuto. Le tre tradizioni concordano nel riconoscere che la categorizzazione e il significato non sono fenomeni binari: vi sono gradi di tipicità (Rosch), gradi di appartenenza al frame (Fillmore), gradi di efficacia delle proiezioni metaforiche (Lakoff). Questa gradualità corrisponde alla struttura metrica della firma informazionale: la distanza di Gromov-Hausdorff tra firme è una quantità continua, e la «gradualità» semantica è semplicemente la gradualità della distanza nello spazio delle firme.

La terza convergenza riguarda la variabilità contestuale. Le tre tradizioni riconoscono che il contenuto semantico varia sistematicamente con il contesto: la profilazione del frame varia con il contesto d'uso (Fillmore), il prototipo attivato varia con il contesto di categorizzazione (Barsalou-Rosch), lo schema di immagine attivato varia con il dominio concettuale (Lakoff). Questa variabilità corrisponde alla variabilità della profilazione della firma informazionale: diversi contesti d'uso attivano diverse proiezioni della firma completa, producendo contenuti contestualmente diversi ma strutturalmente connessi.

X.6,2 — Le differenze metodologiche e il problema della misurabilità

A queste convergenze strutturali si contrappongono alcune differenze metodologiche importanti. La linguistica cognitiva ha sviluppato i suoi strumenti per lo studio del linguaggio naturale: i frame semantici sono

costrutti della semantica lessicale, i prototipi sono oggetti di studio della psicologia cognitiva sperimentale, gli schemi di immagine sono inferenze teoriche sull'organizzazione dell'esperienza pre-linguistica. La teoria della firma informazionale è uno strumento formale quantitativo derivato dalla teoria dell'informazione. La connessione tra i due livelli di analisi — quello qualitativo-linguistico della semantica cognitiva e quello quantitativo-formale della firma informazionale — non è immediata e richiede un'operazionalizzazione.

La connessione più promettente sembra essere quella attraverso i dati di attivazione neurale. I frame semantici di Fillmore possono essere studiati come strutture di co-attivazione nelle neuroimaging studies: attivare il lessema «comprare» dovrebbe attivare simultaneamente, con caratteristiche strutturali specifiche, i nodi neurali corrispondenti agli elementi del frame della transazione commerciale. La struttura di questa co-attivazione — quantificata attraverso le misure di ridondanza, unicità e sinergia della PID applicata ai dati fMRI — dovrebbe riflettere la struttura del frame. I prototipi di Rosch possono essere studiati come attrattori nelle distribuzioni di co-attivazione: gli esemplari prototipici dovrebbero produrre pattern di co-attivazione con alta ridondanza interna (alta coerenza della firma), mentre gli esemplari periferici dovrebbero produrre pattern con più bassa ridondanza. Gli schemi di immagine di Lakoff possono essere studiati come strutture di co-attivazione sensomotoria: l'attivazione di concetti astratti strutturati da schemi di immagine dovrebbe produrre co-attivazione di aree sensorimotorie in pattern coerenti con la struttura dello schema.

Questo programma di ricerca è ambizioso e non ancora pienamente realizzato, ma i suoi elementi componenti hanno già trovato supporto empirico in linee di ricerca convergenti: la ricerca sull'embodied cognition di Barsalou e colleghi, i lavori neuroimaging sulla semantica lessicale di Huth e colleghi, gli studi sulla co-attivazione nella comprensione metaforica di Glucksberg e colleghi. La teoria della ridondanza costitutiva offre il quadro formale unificante entro cui questi risultati dispersi possono essere integrati in una teoria coerente del contenuto semantico incorporato e distribuito.

X.6,3 — La semantica cognitiva come fonte di casi test per la teoria formale

Oltre alla convergenza teorica, la linguistica cognitiva offre alla teoria della ridondanza costitutiva una risorsa metodologica preziosa: un vasto repertorio di casi semantici ben documentati che possono fungere da casi test per le previsioni della teoria formale. I frame semantici di FrameNet, le distribuzioni di tipicità documentate dagli studi di Rosch e colleghi, le strutture delle estensioni metaforiche catalogate da Lakoff e Johnson nel loro lessico delle metafore concettuali — tutto questo costituisce un corpus di fenomeni semantici specifici che una teoria formale del contenuto deve saper spiegare. La firma informazionale, per essere una

teoria genuinamente informativa e non semplicemente un linguaggio di ridescrizione, deve fare previsioni specifiche su questi fenomeni: previsioni sul profilo delle firme delle diverse categorie semantiche, sulle strutture di co-attivazione che diversi frame semantici producono, sulle distanze di Gromov-Hausdorff tra le firme di categorie a livelli di base diversi.

Alcune di queste previsioni sono già verificabili con i dati disponibili. La previsione che i frame semantici corrispondono a strutture di ridondanza ad alta coerenza è consistente con i dati di co-attivazione lessicale provenienti dagli studi di priming semantico: le parole che appartengono allo stesso frame producono effetti di priming reciproci più forti delle parole che appartengono a frame diversi, e la struttura di questi effetti è coerente con la struttura di ridondanza del frame. La previsione che i prototipi corrispondono agli esemplari con firme di massima coerenza interna è consistente con i dati di tipicità gradata: gli esemplari più tipici sono quelli che producono i pattern di co-attivazione più coerenti e stabili tra soggetti diversi. La previsione che gli schemi di immagine corrispondono a strutture di sinergia sensomotoria è consistente con i dati neuroimaging sulla comprensione metaforica: le metafore concettuali attivano aree sensorimotorie in modo sistematico, e la struttura di questa co-attivazione corrisponde alla struttura dello schema di immagine che la metafora proietta.

X.7 — Il problema della polisemia come caso paradigmatico di firma variabile

X.7,1 — La polisemia e le sue teorie

La polisemia — il fenomeno per cui una singola parola ha molti significati correlati tra loro — è uno dei problemi centrali della semantica lessicale ed è anche uno dei fenomeni più illuminanti per la teoria della firma informazionale. La polisemia si distingue dall'omonimia (una parola con molti significati non correlati, come «ramo» nel senso di ramo d'albero e «ramo» nel senso di ramo d'azienda) per il fatto che i significati polisemici sono sistematicamente connessi tra loro: si possono ricavare l'uno dall'altro attraverso trasformazioni semantiche sistematiche come la metonimia, la metafora, la specializzazione contestuale.

Il caso paradigmatico di polisemia sistemamica è quello dei sostantivi che denotano sia un contenitore sia il suo contenuto: «bottiglia», «bicchiere», «scatola» possono denotare sia il contenitore fisico (la bottiglia è rotta) sia il suo contenuto (la bottiglia è finita). Questa alternanza non è arbitraria: è sistematica attraverso tutti i nomi di contenitore della lingua, e produce la medesima ambiguità strutturale. Nel modello dei frame di Fillmore, entrambi i significati sono elementi del frame del contenitore, e la polisemia è la profilazione alternativa di elementi diversi del frame in contesti diversi. Nel modello della firma informazionale, la polisemia è la

proprietà di una parola di avere una firma ricca che include componenti corrispondenti a entrambi i significati, e di poter essere profilata in modo da porre in primo piano le componenti corrispondenti a un significato o all'altro.

X.7,2 — La firma polisemica e la sua struttura

La firma informazionale di una parola polisemica ha una struttura specifica che si distingue sia dalla firma di una parola monosemica (una parola con un solo significato) sia dalla firma di un omonimo (due firme disconnesse associate alla stessa forma fonologica). La firma polisemica ha la forma di una struttura con un nucleo ad alta ridondanza — le componenti informatiche condivise da tutti i significati — circondata da zone periferiche corrispondenti ai significati specifici, collegate al nucleo da relazioni di ridondanza parziale.

Per il caso «bottiglia», la firma include un nucleo ad alta ridondanza corrispondente all'idea di un contenitore cilindrico con gola stretta (componenti visive, tattili, funzionali), e due zone periferiche: una corrispondente al contenitore fisico (con le sue proprietà meccaniche, il suo peso, la sua fragilità) e una corrispondente al contenuto (con le sue proprietà liquide, il suo gusto, il suo volume). Le due zone periferiche sono collegate al nucleo da relazioni di ridondanza parziale — entrambe presuppongono il nucleo — ma non sono direttamente ridondanti tra loro: le proprietà meccaniche del contenitore non sono informative sul gusto del contenuto, e viceversa.

La struttura di sinergia tra le due zone periferiche è bassa ma non nulla: in certi contesti, la qualità del contenitore è correlata con la qualità del contenuto (una bottiglia di champagne di lusso implica champagne di qualità), producendo componenti sinergiche tra le due zone. Questa sinergia contestuale è ciò che rende possibili costruzioni come «questa è una buona bottiglia di vino» (dove sia il contenitore sia il contenuto contribuiscono al giudizio di qualità) che non sono riducibili né all'elogio del contenitore né all'elogio del contenuto.

X.8 — La linguistica cognitiva come campo di validazione empirica indipendente

Il percorso attraverso le tre tradizioni della linguistica cognitiva ha mostrato una convergenza che non era prevedibile a partire dalla sola teoria della firma informazionale: i fenomeni semantici documentati dalla linguistica cognitiva — frame, prototipi, schemi di immagine, strutture radiali, polisemia sistematica, variabilità contestuale — sono riconoscibili, nella loro struttura profonda, come manifestazioni della struttura di ridondanza, unicità e sinergia delle firme informazionali associate ai concetti lessicali. Questa convergenza non è superficiale: è strutturale nel

sensu che si manifesta nelle stesse previsioni specifiche su fenomeni come la gradualità della tipicità, la struttura dei livelli di base, la selettività delle proiezioni metaforiche, la sistematicità della polisemia.

Il contributo di questo capitolo all'argomento complessivo della monografia è duplice. Sul piano teorico, la convergenza con la linguistica cognitiva mostra che la tesi della ridondanza costitutiva non è un'invenzione dei filosofi della mente che lavorano con strumenti formali della teoria dell'informazione: è una tesi che affiora indipendentemente in ogni tradizione di ricerca che si confronta seriamente con la struttura del contenuto semantico nell'uso linguistico reale. Sul piano metodologico, la linguistica cognitiva offre un vasto repertorio di fenomeni empirici ben documentati che possono fungere da banco di prova per le previsioni della teoria formale, rendendo possibile una forma di ricerca interdisciplinare in cui la precisione formale della teoria informatica si combina con la ricchezza fenomenologica della semantica linguistica.

Il capitolo successivo tornerà al piano formale per affrontare una delle previsioni più specifiche della teoria della firma informazionale nella semantica della generalizzazione categoriale: il problema di come un sistema con firma ricca e coerente generalizza a nuove istanze in modo sistematicamente superiore rispetto a sistemi basati su condizioni necessarie e sufficienti o su prototipi isolati.

Capitolo XI — I limiti della teoria: obiezioni, lacune e frontiere aperte

XI.1 — La necessità di un capitolo sui limiti

Una monografia filosofica che pretendesse di non avere difetti, o che relegasse le proprie difficoltà in note a piè di pagina e ammissioni parentetiche nel corpo del testo, tradirebbe il lettore e si esporrebbe a critiche facili. La teoria della ridondanza costitutiva, come ogni proposta teorica ambiziosa che lavori all'interfaccia tra filosofia, matematica e neuroscienze, ha limiti genuini che non sono semplici lacune tecniche destinate a essere colmate da ricerca futura, ma difficoltà strutturali che toccano i fondamenti dell'impresa e che devono essere discusse con la stessa cura con cui si sono esposti i risultati positivi.

Questo capitolo identifica tre famiglie di difficoltà che si ritengono le più serie e le meno facilmente superabili nell'orizzonte attuale della ricerca. La prima famiglia riguarda il problema della scelta dell'obiettivo: la firma informazionale è definita relativa a un obiettivo Θ , ma la scelta di Θ è trattata nel quadro teorico come un dato primitivo, mentre essa stessa

richiede una giustificazione che il quadro non fornisce dall'interno. La seconda famiglia riguarda il problema della circolarità nella definizione della firma di riferimento: la firma di riferimento $\Sigma^*(\Theta)$ — che funge da punto di normazione rispetto a cui si misura l'errore rappresentazionale — è definita in termini dell'obiettivo Θ , ma la determinazione di Θ dipende a sua volta dalla firma di riferimento, creando una circolarità che rischia di rendere la teoria normativa autoconvalidante. La terza famiglia riguarda il problema della computabilità: le misure PID su cui si fonda la firma informazionale hanno un costo computazionale che cresce super-esponenzialmente con il numero delle sorgenti, il che rende la teoria inapplicabile in modo diretto ai sistemi ad alta dimensionalità come il cervello biologico.

L'esposizione di ciascun problema segue un ordine costante: prima la formulazione precisa del problema, poi le risposte disponibili e la loro valutazione, poi la residua difficoltà che rimane dopo le migliori risposte disponibili, e infine le direzioni di ricerca che il problema apre. Non si intende sostenere che la teoria sia insalvabile: si intende sostenere che i limiti indicati richiedono lavoro genuino che non è stato ancora fatto, e che il valore della teoria dipende in parte dal riconoscimento onesto di questi limiti.

XI.2 — Il problema della scelta dell'obiettivo: la T in PID come primitivo ingiustificato

XI.2,1 — L'obiettivo come condizione di possibilità della decomposizione

La decomposizione dell'informazione parziale (PID) che fonda tutta la costruzione della firma informazionale è definita relativa a un obiettivo Y — una variabile aleatoria rispetto alla quale si misura la ridondanza, l'unicità e la sinergia delle sorgenti $X_1, \dots, X_n$. La struttura matematica della decomposizione è impeccabile: dato Y , gli atomi PID sono ben definiti (a meno delle ambiguità nella scelta dell'assioma di non-negatività, di cui si dirà in XI.4), e la firma $\Sigma(\Xi, \Theta)$ che ne risulta è un invariante rigoroso della struttura informazionale del sistema rispetto all'obiettivo. Ma questa impeccabilità tecnica presuppone che l'obiettivo sia dato: la teoria non dice nulla su come si individua l'obiettivo appropriato.

Il problema è reale e non è facilmente liquidabile come questione di applicazione piuttosto che di fondamento teorico. Si consideri il caso del riconoscimento visivo di un volto. Si è argomentato nel Capitolo I che la firma informazionale del sistema visivo rispetto all'obiettivo «volto di Maria» determina il contenuto della rappresentazione. Ma qual è

l'obiettivo? Ci sono almeno cinque candidati distinti, e non è ovvio quale sia quello rilevante per la teoria del contenuto:

- (a) Y = identità del volto (Maria vs. non-Maria): variabile binaria, misura l'informazione sul riconoscimento categoriale;
- (b) Y = configurazione fisiognomica del volto (vettore di caratteristiche geometriche del volto): variabile continua alta-dimensionale, misura l'informazione sul profilo geometrico;
- (c) Y = presenza del volto nella scena visiva: variabile binaria ancora più grezza, misura semplicemente se il volto è visibile o no;
- (d) Y = stato emotivo di Maria (come desunto dall'espressione facciale): variabile categoriale che misura l'informazione sulla valenza emotiva;
- (e) Y = l'insieme dei comportamenti adattativi appropriati rispetto a Maria: variabile funzionale che misura l'informazione sulla risposta comportamentale ottimale.

Ciascuno di questi obiettivi produce una firma informazionale diversa con lo stesso sistema di sorgenti Ξ . La firma rispetto all'obiettivo (a) misura l'informazione di riconoscimento identitario; la firma rispetto all'obiettivo (b) misura l'informazione di rappresentazione geometrica fine; la firma rispetto all'obiettivo (e) misura l'informazione di orientamento comportamentale. Queste firme sono genuinamente diverse — hanno atomi diversi, pesi diversi, strutture topologiche diverse. Se la teoria del contenuto afferma che la firma informazionale determina il contenuto, deve specificare rispetto a quale obiettivo la firma è calcolata, altrimenti la nozione di contenuto è ambigua in modo non accettabile.

XI.2,2 — Risposte disponibili e loro valutazione

Esistono almeno quattro risposte possibili a questo problema, che si esaminano in ordine crescente di plausibilità.

La prima risposta è teleologica: l'obiettivo rilevante è quello che la storia evolutiva ha selezionato come la funzione adattativa primaria del sistema. Questo riavvicina la teoria alla teleosemantica di Millikan, con tutte le difficoltà discusse nel Capitolo III. La risposta è parzialmente corretta — l'evoluzione ha certamente modellato gli obiettivi informativi dei sistemi cognitivi — ma non è sufficiente perché (i) la funzione evolutiva di molti sistemi cognitivi è ambigua o multipla, (ii) i sistemi cognitivi biologici non sono fissi ma si sviluppano nel corso della vita individuale con l'acquisizione di nuovi obiettivi informativi, e (iii) molti contenuti mentali non hanno una storia evolutiva diretta (come i contenuti matematici o i contenuti contrfattuali).

La seconda risposta è funzionalista: l'obiettivo rilevante è quello che meglio spiega il comportamento del sistema — l'obiettivo Y^* tale che la firma $\Sigma(\Xi, Y^*)$ ha il massimo potere predittivo sul comportamento osservato del sistema. Questa risposta ha il vantaggio di essere empiricamente testabile: si possono confrontare le firme rispetto a diversi candidati obiettivi e valutare quale predice meglio il comportamento. Ma ha un problema strutturale: il «comportamento» del sistema è esso stesso un oggetto complesso che richiede descrizione a un certo livello, e la scelta del livello di descrizione del comportamento ripropone il problema della scelta dell'obiettivo a un livello diverso. Se si descrive il comportamento come «l'organismo si avvicina a Maria» (livello grossolano), si otterrà un obiettivo diverso da «l'organismo esibisce il pattern motorio P_{17} » (livello fine-grained). La risposta funzionalista è circolare se l'obiettivo è definito in termini del comportamento e il comportamento è definito in termini dell'obiettivo.

La terza risposta è contestualista: non esiste un unico obiettivo rilevante per un sistema cognitivo in un dato momento, ma una distribuzione di obiettivi che variano con il contesto. La firma informazionale non è definita rispetto a un singolo obiettivo Y ma rispetto a una distribuzione $P(Y)$ di obiettivi possibili, e il contenuto del sistema è la struttura di questa distribuzione di firme. Questa risposta è concettualmente attraente perché cattura la variabilità contestuale del contenuto (discussa nel Capitolo X in relazione alla linguistica cognitiva) ma introduce una complessità tecnica aggiuntiva: definire una firma rispetto a una distribuzione di obiettivi richiede o l'integrazione delle firme su $P(Y)$ (producendo una firma media che può perdere la struttura specifica di ciascun obiettivo) o la considerazione di una famiglia di firme (producendo un oggetto tecnico più ricco ma meno maneggevole).

La quarta risposta — che si ritiene la più soddisfacente pur riconoscendone i limiti — è quella strutturalista: l'obiettivo è determinato dalla struttura del sistema cognitivo stesso, e non deve essere specificato esternamente. Più precisamente, l'obiettivo rilevante per il sistema Ξ rispetto al contenuto Θ è quella variabile Y che massimizza la ricchezza rappresentazionale $\varrho(\Xi, Y)$:

$$\Theta^* = \operatorname{argmax}_Y \varrho(\Xi, Y) = \operatorname{argmax}_Y \kappa(\Sigma(\Xi, Y)) \cdot (1 + \sigma(\Sigma(\Xi, Y)))$$

In questa risposta, l'obiettivo è endogeno: è determinato dalla struttura informazionale del sistema stesso, come quella variabile del mondo che il sistema «cerca di rappresentare» nel senso che la sua architettura è massimamente informativa rispetto ad essa. Questo risponde all'obiezione della scelta arbitraria dell'obiettivo senza fare appello a considerazioni teleologiche esterne, ma introduce una nuova difficoltà: il massimo in questa formula potrebbe non essere unico (potrebbero esserci molti Y con la stessa ricchezza rappresentazionale), e la struttura del sistema potrebbe non determinare un unico obiettivo ma una famiglia di obiettivi egualmente naturali. In questo caso, il contenuto del

sistema non è unico ma è esso stesso una struttura plurale — il che è una conclusione filosoficamente interessante, coerente con la tesi della ridondanza costitutiva, ma che richiede ulteriore elaborazione teorica.

XI.2,3 — La difficoltà residua

Dopo le risposte disponibili, rimane una difficoltà genuina: la risposta strutturalista determina un obiettivo dal sistema ma non garantisce che questo obiettivo corrisponda al contenuto intuitivo della rappresentazione. È possibile che il massimo di $q(\Xi, Y)$ sia raggiunto per un Y che non è quello che un osservatore esterno o il soggetto stesso identificherebbe come il contenuto della rappresentazione.

Consideriamo un caso ipotetico. Un sistema cognitivo elabora informazione proveniente da un volto e da una voce. La ricchezza rappresentazionale potrebbe essere massimizzata dall'obiettivo «identità della persona» (variabile che combina volto e voce), oppure dall'obiettivo «stato emotivo della persona» (che potrebbe essere più informativo rispetto alla struttura combinata delle sorgenti), oppure dall'obiettivo «intenzione comportamentale della persona» (che potrebbe essere l'obiettivo evolutivamente rilevante). Il massimo di q non distingue automaticamente tra questi candidati se la loro ricchezza rappresentazionale è simile, e la teoria non ha risorse per fare la scelta. Questo limite non è banale: indica che la risposta strutturalista determina la forma della firma (come è distribuita l'informazione) senza determinare univocamente il contenuto semantico (di cosa è informativa quella distribuzione). Questa dissociazione tra forma della firma e contenuto semantico è, a giudizio di chi scrive, il limite teorico più serio della teoria della ridondanza costitutiva nel suo stato attuale di sviluppo.

XI.3 — Il problema della circolarità nella definizione della firma di riferimento

XI.3,1 — La firma di riferimento e la sua funzione normativa

La teoria della normatività e dell'errore rappresentazionale sviluppata nel Capitolo VI si basa sulla nozione di firma di riferimento $\Sigma^*(\Theta)$ — la firma informazionale che un sistema cognitivo idealmente calibrato avrebbe rispetto all'obiettivo Θ . La misura di errore è definita come:

$$\text{Err}(\Xi, \Theta) = d_{\text{GH}}(\Sigma(\Xi, \Theta), \Sigma^*(\Theta))$$

e questa misura determina in modo quantitativo la distanza del sistema attuale dalla configurazione di massima accuratezza rappresentazionale. Tutta la teoria della normatività — la distinzione tra rappresentazione corretta e misrappresentazione, la classificazione degli errori, la struttura

della correzione degli errori — si fonda su questa misura, e quindi sulla definizione di $\Sigma^*(\Theta)$.

Il problema è il seguente. $\Sigma^*(\Theta)$ è definita come la firma del sistema cognitivo idealmente calibrato rispetto all'obiettivo Θ . Ma come si determina la firma del sistema «idealmente calibrato»? La risposta implicita nella teoria è: come la firma del sistema che ha massima ridondanza con Θ , ovvero che traccia Θ nel modo più accurato e completo. Ma «tracciare Θ nel modo più accurato» presuppone che Θ sia già identificato come un oggetto nel mondo rispetto al quale si può misurare l'accuratezza della firma. E l'identificazione di Θ dipende, a sua volta, dalla struttura della firma di riferimento: come si è discusso nella sezione precedente, la scelta dell'obiettivo è determinata (nella risposta strutturalista) dalla firma stessa.

La circolarità è dunque la seguente: la firma di riferimento $\Sigma^*(\Theta)$ dipende dall'obiettivo Θ , l'obiettivo Θ è determinato dalla firma stessa (nella risposta strutturalista), e la firma effettiva $\Sigma(\Xi, \Theta)$ è misurata rispetto alla firma di riferimento $\Sigma^*(\Theta)$. Se la catena di dipendenze fosse lineare, non ci sarebbe problema. Ma la risposta strutturalista introduce una dipendenza circolare: Θ dipende da Σ , Σ^* dipende da Θ , e Err dipende da Σ e Σ^* . La struttura circolare mina la pretesa della teoria di fornire una misura indipendente dell'errore rappresentazionale.

XI.3,2 — Tipi di circolarità e la distinzione tra circolarità viciosa e circolarità virtuosa

Prima di valutare la gravità di questa difficoltà, è opportuno distinguere tra due tipi di circolarità. La circolarità viciosa è quella in cui la definizione di un concetto X presuppone il concetto X stesso in modo che non è possibile determinare il valore di X senza già sapere il valore di X . Un classico esempio è la definizione di «buono» come «ciò che le persone buone approvano»: per sapere cosa è buono occorre sapere chi sono le persone buone, ma per sapere chi sono le persone buone occorre già sapere cosa è buono. La circolarità viciosa rende la definizione infondante.

La circolarità virtuosa, al contrario, è quella in cui una definizione circolare converge iterativamente a un punto fisso: partendo da un'approssimazione iniziale di X , si applica ripetutamente la definizione circolare fino a convergere a un valore stabile. I punti fissi delle equazioni auto-referenziali possono essere determinati in modo rigoroso anche quando la definizione è circolare, purché la mappa definita dall'equazione sia una contrazione nello spazio opportuno. Un esempio di circolarità virtuosa è la definizione di equilibrio nei giochi: l'equilibrio di Nash è un profilo di strategie in cui ciascun giocatore gioca la migliore risposta alle strategie degli altri, ma le strategie degli altri dipendono a loro volta dalle aspettative su cosa ciascun giocatore farà. La circolarità è reale ma non viciosa: esistono teoremi di punto fisso (Kakutani, Nash) che

garantiscono l'esistenza di un equilibrio senza richiedere una definizione non-circolare delle strategie.

La questione è dunque se la circolarità nella definizione della firma di riferimento sia viciosa o virtuosa. La risposta dipende dalla struttura della mappa circolare. Se la mappa

$$F : \Sigma \mapsto \Sigma^* (\Theta(\Sigma))$$

dove $\Theta(\Sigma)$ è l'obiettivo determinato dalla firma Σ e $\Sigma^*(\Theta)$ è la firma di riferimento per quell'obiettivo, è una contrazione nello spazio delle firme dotato della metrica d_{GH} , allora la circolarità è virtuosa: esiste un punto fisso $\Sigma^{**} = F(\Sigma^{**})$, e questo punto fisso è la firma di riferimento auto-consistente che soddisfa simultaneamente il requisito che l'obiettivo sia determinato dalla firma e il requisito che la firma sia calibrata rispetto all'obiettivo.

Non si è in grado, nel presente stato della teoria, di dimostrare che F è una contrazione in generale. Esistono ragioni plausibili per credere che F sia una contrazione in condizioni ragionevoli — condizioni in cui il sistema cognitivo ha struttura informazionale stabile e l'obiettivo non varia drasticamente con piccole perturbazioni della firma — ma una dimostrazione rigorosa richiederebbe ipotesi sul legame tra la struttura della firma e la struttura dell'obiettivo che non sono state formalizzate.

XI.3,3 — Una versione deflazionista della firma di riferimento

Una risposta parziale al problema della circolarità è quella deflazionista: abbandonare la pretesa di definire la firma di riferimento in modo assoluto e definirla invece in modo relativo a una popolazione di riferimento di sistemi cognitivi. Nella versione deflazionista:

$$\Sigma^*(\Theta, \text{Pop}) = \operatorname{argmin}_{\{\Sigma\}} \Sigma_{\{i \in \text{Pop}\}} d_{GH}(\Sigma, \Sigma_i(\Theta))$$

dove $\text{Pop} = \{\Xi_1, \dots, \Xi_N\}$ è la popolazione di sistemi di riferimento (ad esempio, tutti gli adulti neurologicamente tipici di una data cultura) e $\Sigma_i(\Theta)$ è la firma del sistema Ξ_i rispetto all'obiettivo Θ . La firma di riferimento deflazionista è la media di Fréchet delle firme della popolazione — il punto dello spazio delle firme che minimizza la somma delle distanze al quadrato dalle firme dei sistemi della popolazione. Questa definizione non è circolare: richiede solo le firme effettive dei sistemi della popolazione, non una firma di riferimento assoluta.

Il prezzo della risposta deflazionista è la relatività della normatività. L'errore rappresentazionale non è più una nozione assoluta — la distanza da una firma di riferimento oggettiva — ma è una nozione relativa alla popolazione di riferimento scelta. Questo ha implicazioni per la psicopatologia (l'errore rappresentazionale diventa relativo alla norma culturale, non a una struttura oggettiva del mondo), per la filosofia della

scienza (il cambiamento concettuale nella scienza non è più un avvicinamento a una firma di riferimento assoluta ma una ri-definizione della popolazione di riferimento), e per la teoria dell'AI (l'allineamento dei sistemi artificiali non è calibrazione rispetto a una firma ideale ma approssimazione alla firma media umana, che è essa stessa variabile culturalmente e storicamente).

Queste implicazioni non sono necessariamente negative: molte di esse catturano intuizioni filosofiche plausibili sulla natura contestuale e relativa della normatività rappresentazionale. Ma non eliminano completamente il problema della circolarità: la scelta della popolazione di riferimento Pop è essa stessa soggetta a giustificazione, e la giustificazione della scelta di Pop presuppone criteri normativi che la teoria deflazionista non può fornire senza circolarità.

XI.3,4 — La difficoltà residua e le sue implicazioni teoriche

La difficoltà che rimane dopo le risposte disponibili è la seguente: la teoria della ridondanza costitutiva fornisce una misura dell'errore rappresentazionale che è internamente coerente e formalmente precisa, ma non ha i mezzi per giustificare indipendentemente il punto di normazione rispetto a cui l'errore è misurato. Questa è una limitazione che la teoria condivide con molte teorie naturalistiche della normatività: anche la teleosemantica di Millikan e la teoria dretsiana non riescono a giustificare completamente il punto di normazione senza fare appello a circolarità o a primitivi normativi irriducibili. La critica che la teoria della ridondanza costitutiva riceve su questo fronte è dunque una versione di una critica generale che colpisce ogni teoria naturalistica del contenuto che aspira a spiegare la normatività.

Questa circostanza non esime dalla critica, ma la relativizza: il problema della circolarità nella firma di riferimento è reale, ma non è peculiare alla teoria della ridondanza costitutiva, ed è plausibile che qualsiasi teoria naturalistica del contenuto con pretese normative debba affrontare una difficoltà strutturalmente analoga. Il contributo specifico della teoria della ridondanza costitutiva è che la struttura formale della firma informazionale rende il problema più preciso e potenzialmente più trattabile: la questione del punto di normazione può essere formulata come la questione dell'esistenza di un punto fisso della mappa F definita sopra, e questa è una questione matematica rigorosa che ammette in linea di principio risposte rigorose.

XI.4 — Il problema della computabilità delle misure PID in alta dimensionalità

XI.4,1 — La crescita combinatoria del reticolo degli atomi

Il problema computazionale della decomposizione PID è ben noto agli esperti del campo e costituisce uno dei principali ostacoli all'applicazione empirica della teoria. Per una decomposizione con n sorgenti, il reticolo degli atomi ha una dimensione che cresce in modo super-esponenziale con n . Il numero di antichain del reticolo degli antichain dell'insieme potenza di n elementi — che è il numero di atomi nella decomposizione PID per n sorgenti — è dato dalla sequenza di Dedekind $D(n)$, i cui valori sono:

$$D(0) = 1 \quad D(1) = 3 \quad D(2) = 6 \quad D(3) = 20 \quad D(4) = 168 \quad D(5) = 7581 \quad D(6) = 7828354 \quad D(7) = 2414682040998$$

La crescita è dunque straordinariamente rapida: da $n=6$ a $n=7$ il numero di atomi cresce di un fattore di circa 300. Per $n=10$ (un sistema con 10 sorgenti), il numero di atomi è dell'ordine di 10^{23} , del tutto inaccessibile al calcolo esplicito. Il cervello biologico ha dell'ordine di 10^{11} neuroni, e anche se si considerano moduli funzionali (aree corticali, reti funzionali) piuttosto che neuroni individuali, il numero di sorgenti rilevanti per la firma informazionale di qualsiasi contenuto complesso è di gran lunga superiore a 10, rendendo il calcolo esatto della decomposizione PID computazionalmente impossibile.

Questo limite non è solo pratico ma tocca il fondamento teorico: se la firma informazionale è definita esattamente attraverso la decomposizione PID, e la decomposizione PID è computazionalmente inaccessibile per i sistemi biologicamente rilevanti, allora la firma è un oggetto teorico la cui istanziazione concreta non è misurabile con metodi attualmente disponibili.

XI.4,2 — Il problema dell'unicità dell'assioma di non-negatività

Prima di discutere le risposte al problema computazionale, è opportuno affrontare un problema logicamente anteriore che riguarda la definizione stessa della PID. La decomposizione PID non è ancora univocamente definita nella letteratura, perché diverse proposte per l'atomizzazione delle componenti di ridondanza, unicità e sinergia soddisfano assiomi diversi e producono decomposizioni diverse. Le principali proposte in competizione sono:

$I_{\min}$ di Williams e Beer (2010): definisce la ridondanza come il minimo delle informazioni mutue bilaterali. Soddisfa l'assioma di non-negatività (tutti gli atomi sono non-negativi) per $n=2$ sorgenti, ma per $n \geq 3$ può produrre atomi negativi in certi scenari.

I_{ccs} di Bertschinger et al. (2014): definisce la ridondanza attraverso la massimizzazione dell'entropia condizionale dei margini. Soddisfa la non-negatività per tutti n , ma ha la proprietà controintuitiva che la ridondanza tra sorgenti statisticamente indipendenti può essere positiva.

I_BROJA di Bertschinger et al. (2014): usa una soluzione di ottimizzazione convessa che produce atomi non-negativi con proprietà assiomatiche soddisfacenti, ma è computazionalmente più onerosa e ha problemi di unicità in alcuni casi degeneri.

I_{sx} di Griffith e Koch (2014): definisce la sinergia come la differenza tra informazione mutua totale e somma delle informazioni mutue parziali massimali, con proprietà di non-negatività ma a spese di una ridondanza che non cattura tutte le intuizioni sulla condivisione di informazione.

La molteplicità di queste proposte indica che la PID non è ancora una teoria consolidata: è un programma di ricerca attivo in cui la definizione «corretta» degli atomi non è stata determinata definitivamente. La firma informazionale definita in questa monografia presuppone implicitamente una scelta tra queste proposte, e risultati diversi potrebbero essere ottenuti con scelte diverse. Questa dipendenza dalla scelta della misura PID è un problema reale per la teoria: affermare che la firma informazionale determina il contenuto richiede che la firma sia ben definita, ma la firma è ben definita solo a meno di una scelta tra misure PID alternative che la teoria non giustifica.

XI.4,3 — Risposte al problema computazionale: approssimazioni e misure surrogate

Diverse strategie sono state proposte per rendere la PID computazionalmente trattabile in sistemi ad alta dimensionalità. Si valutano le principali, con particolare attenzione alle implicazioni teoriche per la firma informazionale.

La prima strategia è la decomposizione gerarchica: invece di calcolare la PID per tutte le n sorgenti simultaneamente, si calcola per sottosistemi di dimensione ridotta (tipicamente coppie o triple) e si aggregano i risultati attraverso una procedura gerarchica. Questo approccio riduce la complessità computazionale da $O(D(n))$ a $O(n^2 \cdot D(2)) = O(n^2)$ per le decomposizioni a coppie, che è polinomiale in n e quindi computazionalmente accessibile anche per sistemi grandi. Il prezzo è la perdita delle componenti sinergiche di ordine superiore: la decomposizione gerarchica a coppie può catturare ridondanza e sinergia binarie, ma non la sinergia di ordine k per $k > 2$. Se le componenti sinergiche di ordine superiore sono semanticamente rilevanti — come si è argomentato nel Capitolo III nel contesto del fallimento della composizionalità — allora l'approssimazione gerarchica perde precisamente le componenti più teoricamente interessanti.

La seconda strategia è l'uso di misure surrogate che approssimano la firma PID senza calcolarne la decomposizione esatta. Una misura surrogata promettente è l'O-information di Rosas e colleghi (2019), definita come:

$$\Omega(X_1, \dots, X) = (n-2) H(X_1, \dots, X) + \sum_i H(X_i) - \sum_i H(X_1, \dots, X_i, \dots, X)$$

dove X_i indica che X_i è omessa. Questa misura cattura il bilanciamento tra ridondanza e sinergia a livello globale: $\Omega > 0$ indica che il sistema è prevalentemente ridondante, $\Omega < 0$ che è prevalentemente sinergico, $\Omega = 0$ che ridondanza e sinergia si bilanciano. La O-information è computazionalmente accessibile (richiede solo il calcolo di entropie marginali e congiunte, che sono $O(2^n)$ ma con $n \leq 20-30$ sono calcolabili con metodi di campionamento) e ha una struttura assiomatica ragionevole. Tuttavia, essendo una misura scalare del bilanciamento globale, non cattura la struttura dettagliata della firma: due sistemi con la stessa O-information possono avere firme profondamente diverse, con distribuzioni diverse tra atomi di ridondanza, unicità e sinergia.

La terza strategia è quella delle misure di sinergia di coppia-livello, sviluppate da Shorten e Nitzan (2021) e da Varley e colleghi. Queste misure calcolano la sinergia tra coppie di sorgenti rispetto all'obiettivo e le aggregano per approssimare la struttura della firma:

$$\tilde{S}(X_1, \dots, X; Y) \approx \sum_{\{i < j\}} S(X_i, X_j; Y)$$

Questa approssimazione è computazionalmente accessibile ($O(n^2)$ calcoli di sinergia binaria) e cattura la struttura delle dipendenze sinergiche binarie, perdendo però le sinergie di ordine superiore esattamente come la decomposizione gerarchica.

La quarta strategia, la più recente e promettente, è quella basata sull'apprendimento automatico per stimare le componenti PID da grandi dataset empirici senza calcolare la decomposizione esatta. Metodi basati su variational autoencoders e su reti neurali per la stima di informazione mutua (come MINE di Belghazi e colleghi) permettono di stimare componenti della firma PID da dati empirici ad alta dimensionalità senza calcolare esplicitamente le distribuzioni congiunte. Questi metodi hanno il vantaggio di scalare a sistemi con decine o centinaia di sorgenti, ma hanno limiti di garanzie di convergenza e possono essere distorte in direzioni sistematiche da bias nell'architettura della rete neurale usata per la stima.

XI.4,4 — Il problema teorico più profondo: la PID come idealizzazione

Le strategie di approssimazione discusse sopra trattano il problema computazionale come un problema di calcolo: come si calcola la firma PID in modo praticamente accessibile? Ma vi è un problema teorico più profondo che va al di là del calcolo: la firma PID esatta è un oggetto ideale che presuppone la conoscenza della distribuzione di probabilità congiunta $P(X_1, \dots, X, Y)$ con precisione arbitraria, e questa conoscenza non è mai disponibile in un sistema biologico reale.

In un sistema cognitivo biologico, la «distribuzione di probabilità» che definisce le relazioni tra le diverse sorgenti di rappresentazione non è un oggetto dato: è una struttura emergente dall'esperienza accumulata del sistema, variabile nel tempo, approssimata e parziale. Il cervello non «conosce» la distribuzione congiunta delle sue proprie attivazioni: implementa comportamenti che sono coerenti con una certa struttura di dipendenze statistiche, ma questa struttura non è mai completamente accessibile al sistema stesso né all'osservatore esterno con strumenti di misura finiti. La PID è dunque, nella sua applicazione ai sistemi biologici, una idealizzazione che presuppone una precisione nella conoscenza della distribuzione che non è realizzabile in principio.

Questa circostanza non rende inutile la teoria: molte teorie fisiche e matematiche lavorano con idealizzazioni che non sono realizzabili in principio (gas perfetti, punti materiali, oscillatori armonici) ma che catturano strutture reali dei sistemi fisici in modo approssimativo. Il punto rilevante è che la teoria della ridondanza costitutiva, nella sua formulazione attuale, non specifica esplicitamente in quale senso la firma PID approssima la struttura informazionale reale del sistema cognitivo, e non fornisce garanzie di convergenza: non è stabilito quanto una firma calcolata su dati finiti (con inevitabile errore di stima) si avvicini alla firma teorica calcolata sulla distribuzione esatta. Questa lacuna è riempibile in linea di principio — attraverso teorie dell'inferenza statistica per le misure PID — ma non è stata ancora colmata in modo soddisfacente.

XI.4,5 — Implicazioni per l'applicabilità della teoria

Le tre lacune individuate nella sezione XI.4 — il costo computazionale super-esponenziale, la mancanza di una definizione PID univocamente accettata, e il problema dell'idealizzazione — si combinano per limitare l'applicabilità della teoria della ridondanza costitutiva ai sistemi biologici reali in modo significativo.

Nel breve periodo, la teoria è applicabile in modo rigoroso solo a sistemi a bassa dimensionalità ($n \leq 4-5$ sorgenti) e a modelli computazionali in cui le distribuzioni di probabilità sono note esattamente. Per sistemi di dimensionalità maggiore, come le reti neurali corticali, la teoria può essere applicata solo attraverso approssimazioni che comportano perdite di informazione teoricamente rilevanti.

Nel medio periodo, è ragionevole attendersi progressi nei metodi di approssimazione della PID che permettano applicazioni a sistemi di dimensionalità moderata ($n \leq 20-30$) con garanzie di convergenza. Questi progressi sono già in corso nella letteratura della neuroscienza computazionale.

Nel lungo periodo, la teoria richiede o un fondamento alternativo che non si basi sulla PID esatta ma su approssimazioni con garanzie teoriche,

oppure una revisione concettuale che sostituisca la PID con un formalismo di complessità polinomiale che preservi le proprietà teoriche rilevanti. Questa revisione non è banale: la complessità super-esponenziale della PID riflette la complessità intrinseca del problema di misurare le dipendenze informatiche di ordine superiore, e non è ovvio che esista un formalismo più semplice che catturi le stesse strutture.

XI.5 — Ulteriori limitazioni: confini e problemi aperti

XI.5,1 — Il problema dell'intenzionalità di ordine superiore

La teoria della ridondanza costitutiva è stata sviluppata principalmente per il caso delle rappresentazioni di primo ordine: rappresentazioni di oggetti, proprietà e relazioni del mondo fisico e sociale. Non si è affrontato il caso delle rappresentazioni di ordine superiore — le rappresentazioni di rappresentazioni, le credenze sulle credenze, le intenzioni sulle intenzioni. Queste strutture di ordine superiore — che formano il nucleo dell'intenzionalità riflessiva caratteristica della cognizione umana adulta — pongono sfide specifiche alla teoria.

Una credenza di secondo ordine del tipo «Credo che Paolo creda che piova» ha una struttura informazionale che include, come componente, la firma informazionale della credenza di primo ordine di Paolo. La firma della credenza di secondo ordine è dunque una firma di una firma, il che solleva la questione di come si calcola la PID su oggetti che sono essi stessi strutture informazionali complesse. La questione non è semplicemente tecnica: tocca il problema di se e come la teoria della ridondanza costitutiva possa essere estesa all'intenzionalità riflessiva senza perdere le proprietà formali che la rendono interessante.

XI.5,2 — Il problema del soggetto e della prospettiva

La teoria della ridondanza costitutiva tratta il sistema cognitivo Ξ come un oggetto terzo-personale — un sistema osservabile dall'esterno, di cui si può calcolare la firma informazionale. Ma il contenuto mentale, almeno nella concezione fenomenologica tradizionale, ha una struttura essenzialmente prima-personale: è il contenuto «per il soggetto», con una struttura prospettica che non è riducibile alla struttura osservabile dall'esterno. La teoria della firma informazionale non ha, in modo esplicito, le risorse per trattare questa dimensione prima-personale del contenuto.

Questo non è necessariamente un limite insuperabile: è possibile che la struttura prospettica del contenuto prima-personale sia codificata nella firma informazionale attraverso la struttura specifica dei sistemi

interoceptivi e somatosensoriali che contribuiscono alle rappresentazioni del sé. Ma la connessione tra la firma informazionale e la struttura prospettica del contenuto fenomenologico non è stata elaborata in questa monografia, e rimane una lacuna significativa che richiederebbe un lavoro ulteriore all'interfaccia con la filosofia della mente fenomenologica.

XI.5,3 — Il problema della grounding per i sistemi artificiali

Nel Capitolo VIII si è discusso il problema del grounding causale per i sistemi di intelligenza artificiale: la tesi che le firme informazionali dei sistemi artificiali addestrati su corpus siano «simulate» piuttosto che «genuine» perché non sono ancorate a esperienze causali dirette con il mondo. Ma la distinzione tra firma simulata e firma genuina non è stata resa sufficientemente precisa.

Si è introdotta la distanza di ancoraggio causale Δ_{causal} come misura della discrepanza tra la firma effettiva del sistema e la firma che avrebbe un sistema con canali causali genuinamente indipendenti. Ma questa misura presuppone che si possa determinare cosa significhi «genuinamente indipendente» per i canali causali, il che richiede una teoria del grounding causale che non è stata sviluppata in modo autonomo. È possibile che un sistema artificiale con struttura di dati sufficientemente ricca e diversificata abbia una Δ_{causal} piccola pur non avendo mai interagito direttamente con il mondo fisico, il che farebbe emergere la distinzione tra firma simulata e firma genuina come una distinzione di grado piuttosto che di tipo. Questa possibilità è teoricamente interessante ma rimane da esplorare in dettaglio.

XI.6 — Una valutazione complessiva

Il bilancio di questo capitolo non deve essere letto come una confutazione della teoria della ridondanza costitutiva, ma come una mappa delle sue frontiere aperte. Ogni teoria filosofica ambiziosa che lavori all'interfaccia tra discipline diverse ha un territorio ben mappato — dove i risultati sono solidi e le argomentazioni stringenti — e una periferia meno definita dove le questioni rimangono aperte. Il territorio ben mappato della teoria della ridondanza costitutiva include: la critica delle teorie standard del contenuto e la diagnosi della presupposizione di unicità come loro punto cieco comune (Capitolo III); la costruzione formale della firma informazionale come reticolo degli atomi PID e le sue proprietà matematiche (Capitolo IV); la metrica sullo spazio dei contenuti e gli invarianti topologici (Capitoli V e V.6 espanso); la teoria dell'errore rappresentazionale e della normatività emergente come distanza dalla firma di riferimento (Capitolo VI); le convergenze con la linguistica cognitiva (Capitolo X). La periferia meno definita include: la giustificazione della scelta dell'obiettivo, la fondazione non-circolare della

normatività, l'accessibilità computazionale per sistemi ad alta dimensionalità, l'intenzionalità di ordine superiore, e il trattamento della prospettiva prima-personale.

Che la periferia sia meno definita del centro non è una critica ma una caratteristica strutturale di qualsiasi programma di ricerca vitale. Il valore di una teoria non si misura soltanto dai problemi che risolve ma anche dai problemi che articola in modo sufficientemente preciso da renderli suscettibili di ricerca futura. Su questo criterio, la teoria della ridondanza costitutiva contribuisce in modo genuino: i problemi del punto XI.2, XI.3 e XI.4 non erano formulabili in questi termini precisi prima che la teoria li inquadrasse, e la loro formulazione precisa è già un contributo al progresso filosofico anche indipendentemente dalle soluzioni che si riusciranno o meno a trovare.

XI.A — Le misure PID alternative e le loro implicazioni per la firma

Come si è osservato nella sezione XI.4,2, la decomposizione PID non è univocamente definita nella letteratura attuale. In questa appendice si precisa in quale modo la scelta della misura PID influisce sulla firma informazionale e sulla teoria del contenuto che su di essa si basa.

La misura $I_{\min}$ di Williams e Beer è definita come:

$$R^{\{WB\}}(\alpha; Y) = \min_{\{ \text{source} \in \alpha \}} I(\text{source}; Y \mid \text{other sources})$$

dove il minimo è preso su tutte le sorgenti contenute nell'antichain α (interpretato come insieme di sorgenti che «competono» per fornire la stessa informazione su Y). Questa misura ha la proprietà di simmetria (invariante sotto permutazioni delle sorgenti) e di monotonicità rispetto all'ordine del reticolo, ma può produrre atomi negativi per $n \geq 3$. Un atomo negativo nella firma significa che la componente corrispondente «sottrae» informazione anziché aggiungerla, il che può accadere quando la ridondanza di ordine superiore è inferiore alla ridondanza di ordine inferiore in un modo che viola la monotonicità attesa.

La misura I_{BROJA} di Bertschinger et al. è definita attraverso il problema di ottimizzazione:

$$R^{\{BROJA\}}(X_1, X_2; Y) = \max_{\{Q \in \Delta(P)\}} I_Q(X_1; Y) + I_Q(X_2; Y) - I_Q(X_1, X_2; Y)$$

dove $\Delta(P)$ è l'insieme delle distribuzioni di probabilità che hanno le stesse distribuzioni marginali di coppia di P (bivariate). Questa misura è sempre non-negativa e ha buone proprietà assiomatiche per $n=2$, ma la sua generalizzazione a $n>2$ non è univoca.

L'implicazione per la firma informazionale è che la struttura del reticolo degli atomi — e dunque gli invarianti topologici — dipende dalla scelta della misura PID. Sistemi che la firma $I_{\min}$ classifica come aventi alta sinergia possono essere classificati come aventi bassa sinergia dalla firma I_{BROJA} , se i loro profili di dipendenza cadono in una zona in cui le due misure divergono. Questa dipendenza dalla misura è una forma di ambiguità teorica che la teoria della ridondanza costitutiva non può risolvere dall'interno: richiede o un argomento indipendente per la scelta della misura, o la formulazione di una versione della teoria robusta rispetto alla scelta della misura (invariante rispetto alle diverse decomposizioni PID accettabili).

Una direzione promettente è quella di un'assiomatizzazione della firma informazionale che non dipenda da una scelta specifica di misura PID ma specifichi direttamente le proprietà che la firma deve soddisfare — analogamente a come l'assiomatizzazione di Shannon dell'entropia non presuppone una formula specifica ma deriva la formula dai suoi assiomi. Un insieme di assiomi plausibili per la firma informazionale include: non-negatività di tutti gli atomi, monotonicità rispetto all'ordine del reticolo, simmetria rispetto alle permutazioni delle sorgenti, self-redundancy ($I(X; Y)$ per una singola sorgente X è un caso speciale di ridondanza pura), e continuità rispetto alle perturbazioni della distribuzione di probabilità congiunta. Non è noto se questi assiomi determinino univocamente una misura PID, e la questione dell'unicità dell'assiomatizzazione è un problema aperto della teoria dell'informazione che ha implicazioni dirette per la fondazione della teoria della ridondanza costitutiva.

Appendice Matematica — Sommario esteso delle definizioni formali principali

Questa appendice raccoglie in forma sistematica tutte le definizioni formali introdotte nel testo, completandole con derivazioni e commenti che nel corpo del testo erano stati omessi per non appesantire l'esposizione. Si seguono i riferimenti ai paragrafi del testo principale in cui ciascuna nozione è stata introdotta e discussa nel suo contesto filosofico.

A.1 — Entropia e informazione mutua: le nozioni primitive

Prima di enunciare le definizioni specifiche alla teoria della ridondanza costitutiva, conviene richiamare le nozioni primitive della teoria dell'informazione da cui il quadro formale dipende.

Sia X una variabile aleatoria con supporto finito $\mathcal{X}$ e distribuzione $P(X = x) = p(x)$. L'entropia di Shannon di X è:

$$H(X) = -\sum_{\{x \in \mathcal{X}\}} p(x) \cdot \log_2 p(x)$$

con la convenzione che $0 \cdot \log_2 0 = 0$. L'entropia è non-negativa, $H(X) \geq 0$, e misura l'incertezza media sulla variabile X : è massima (uguale a $\log_2 |\mathcal{X}|$ bit) quando X è uniformemente distribuita su $\mathcal{X}$, ed è zero quando X è deterministica.

L'entropia congiunta di due variabili X e Y è:

$$H(X, Y) = -\sum_{\{x,y\}} p(x,y) \cdot \log_2 p(x,y)$$

L'entropia condizionale di X dato Y è:

$$H(X|Y) = -\sum_{\{x,y\}} p(x,y) \cdot \log_2 p(x|y) = H(X,Y) - H(Y)$$

L'informazione mutua tra X e Y è:

$$I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$$

L'informazione mutua è non-negativa e simmetrica: $I(X; Y) = I(Y; X) \geq 0$. È zero se e solo se X e Y sono statisticamente indipendenti.

L'informazione mutua condizionale tra X e Y dato Z è:

$$I(X; Y|Z) = H(X|Z) - H(X|Y,Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z)$$

A.2 — La co-informazione (interazione di ordine 3) e la sua relazione con ridondanza e sinergia

Per tre variabili X_1, X_2, Y la co-informazione è definita come:

$$C(X_1; X_2; Y) = I(X_1; Y) - I(X_1; Y|X_2) = I(X_1; Y) + I(X_2; Y) - I(X_1, X_2; Y)$$

A differenza dell'entropia e dell'informazione mutua, la co-informazione può essere negativa. Un valore positivo di C indica ridondanza netta: le due sorgenti condividono informazione sull'obiettivo (sapere X_1 riduce l'informazione aggiuntiva di X_2 su Y). Un valore negativo indica sinergia

netta: le due sorgenti si complementano nell'informare sull'obiettivo (sapere X_1 aumenta l'informazione aggiuntiva di X_2 su Y).

La relazione tra co-informazione e atomi PID per $n = 2$ è:

$$C(X_1; X_2; Y) = \mathcal{R}(\{1\} \wedge \{2\}; Y) - \text{Syn}(\{1,2\}; Y)$$

dove $\mathcal{R}(\{1\} \wedge \{2\}; Y)$ è la ridondanza bilaterale e $\text{Syn}(\{1,2\}; Y)$ è la sinergia bilaterale. La co-informazione è dunque la differenza tra ridondanza e sinergia: misura il "surplus di ridondanza" rispetto alla sinergia.

La co-informazione totale tra n variabili è definita ricorsivamente:

$$C(X_1; X_2; \dots; X_n; Y) = C(X_1; \dots; X_{n-1}; Y) - C(X_1; \dots; X_{n-1}; Y | X_n)$$

ed è una misura dell'interazione statistica di ordine n tra le variabili.

A.3 — Il reticolo degli atomi informativi della PID

Per n sorgenti $S = \{X_1, \dots, X_n\}$ e un obiettivo Y , sia $\mathcal{A}$ l'insieme delle anti-chain del reticolo delle partizioni di $\{1, \dots, n\}$ nel senso del reticolo degli ordini di Dedekind. Un'anti-chain $\alpha \in \mathcal{A}$ è un insieme di sottoinsiemi non vuoti di $\{1, \dots, n\}$ tale che nessun elemento di α è sottoinsieme di un altro elemento di α .

Il reticolo delle anti-chain è dotato dell'ordine parziale: $\alpha \leq \beta$ (α è "più ridondante di" β) se e solo se per ogni $A \in \alpha$ esiste un $B \in \beta$ tale che $A \subseteq B$.

L'atomo informativo $\mathcal{R}(\alpha; Y)$ per $\alpha \in \mathcal{A}$ è la componente dell'informazione mutua $I(S; Y)$ attribuibile univocamente alla struttura di ridondanza caratterizzata da α . La decomposizione è:

$$I(S; Y) = \sum_{\alpha \in \mathcal{A}} \mathcal{R}(\alpha; Y)$$

con la condizione di consistenza marginale: per ogni $A \subseteq \{1, \dots, n\}$,

$$I(X_A; Y) = \sum_{\alpha \leq \{A\}} \mathcal{R}(\alpha; Y)$$

dove $\{A\}$ denota l'anti-chain singleton $\{A\}$ e la somma è presa su tutte le anti-chain α che "raffinano" la struttura di A .

Per $n = 2$, $\mathcal{A}_2$ ha esattamente quattro elementi: $\{1\} \wedge \{2\}$ (ridondanza), $\{1\}$ (unicità di 1), $\{2\}$ (unicità di 2), $\{1,2\}$ (sinergia). La decomposizione è:

$$I(\{X_1, X_2\}; Y) = \mathcal{R}(\{1\} \wedge \{2\}; Y) + U_1(Y) + U_2(Y) + \text{Syn}(\{1, 2\}; Y)$$

con le relazioni marginali:

$$I(X_1; Y) = \mathcal{R}(\{1\} \wedge \{2\}; Y) + U_1(Y) \quad I(X_2; Y) = \mathcal{R}(\{1\} \wedge \{2\}; Y) + U_2(Y)$$

Da cui si ricava:

$$\mathcal{R}(\{1\} \wedge \{2\}; Y) = I(X_1; Y) + I(X_2; Y) - [\mathcal{R}(\{1\} \wedge \{2\}; Y) + U_1 + U_2 + \text{Syn}] + \text{Syn} - U_1 - U_2$$

che richiede una definizione indipendente di uno degli atomi per determinare tutti gli altri. La definizione originale di Williams e Beer usa:

$$\mathcal{R}_{\text{WB}}(\{1\} \wedge \{2\}; Y) = \sum_{-Y} p(y) \cdot \min_{i \in \{1, 2\}} I(X_i; Y=y)$$

dove $I(X_i; Y=y)$ è l'informazione puntuale di X_i rispetto all'evento $\{Y=y\}$:

$$I(X_i; Y=y) = \sum_{x_i} p(x_i | y) \cdot \log_2(p(x_i | y) / p(x_i))$$

A.4 — La firma informazionale: definizione completa

Sia $\Xi = (\xi_1, \dots, \xi_n)$ un vettore di rappresentazioni ridondanti in un agente cognitivo A , e sia Θ la variabile che denota il contenuto dell'oggetto o della proprietà del mondo che il sistema rappresenta. Si suppone che Ξ e Θ siano variabili aleatorie con distribuzione congiunta $P(\Xi, \Theta)$ che riflette la storia delle interazioni tra il sistema e l'ambiente.

La firma informazionale di Ξ rispetto a Θ è la tripla:

$$\Sigma(\Xi, \Theta) = (\mathcal{R}(\Xi, \Theta), \kappa(\Xi, \Theta), \sigma(\Xi, \Theta))$$

dove i tre componenti sono:

Il reticolo degli atomi $\mathcal{R}(\Xi, \Theta)$ è l'insieme pesato $\{\mathcal{R}(\alpha; \Theta) : \alpha \in \mathcal{A}\}$ degli atomi della decomposizione PID di $I(\xi_1, \dots, \xi_n; \Theta)$. Esso è un elemento dello spazio $\mathbb{R}^{|\mathcal{A}|}$ dei vettori di atomi.

Il coefficiente di coerenza $\kappa(\Xi, \Theta)$ è:

$$\kappa(\Xi, \Theta) = [\sum_i I(\xi_i; \Theta) - I(\xi_1, \dots, \xi_n; \Theta)] / [(1/n) \cdot \sum_i I(\xi_i; \Theta)]$$

Questa quantità è pari a $1 - I_{\text{unique_total}} / I_{\text{avg}}$, dove $I_{\text{unique_total}} = \sum_i U_i(\Theta)$ è la somma delle informazioni uniche e I_{avg} è l'informazione media per sorgente. Il coefficiente κ misura la "coesione" del sistema: è alto quando le sorgenti si rinforzano reciprocamente (alta

ridondanza, bassa unicità relativa), e basso quando ciascuna sorgente porta informazione prevalentemente propria.

Il rapporto sinergico $\sigma(\Xi, \Theta)$ è:

$$\sigma(\Xi, \Theta) = \text{Syn}(\{1, \dots, n\}; \Theta) / I(\xi_1, \dots, \xi; \Theta)$$

dove $\text{Syn}(\{1, \dots, n\}; \Theta)$ è la sinergia di ordine n . Il rapporto σ misura la frazione dell'informazione congiunta di natura sinergica: è 0 quando tutta l'informazione congiunta è decomponibile, e tende a 1 quando l'informazione è quasi interamente sinergica.

Una proprietà importante della firma è l'invarianza parziale rispetto alle trasformazioni invertibili locali delle componenti: se $T = (T_1, \dots, T)$ con ciascun T_i invertibile, allora $I(T_1(\xi_1), \dots, T(\xi); \Theta) = I(\xi_1, \dots, \xi; \Theta)$, ma la struttura degli atomi PID si trasforma in modo non banale perché le dipendenze condizionali cambiano. Questo è il senso in cui la firma è sensibile alla struttura delle rappresentazioni, non solo alla quantità totale di informazione che portano.

A.5 — La distanza di Gromov-Hausdorff tra firme

La distanza tra due firme $\Sigma_A = (\mathcal{R}_A, \kappa_A, \sigma_A)$ e $\Sigma_B = (\mathcal{R}_B, \kappa_B, \sigma_B)$ rispetto al medesimo obiettivo Θ è definita come segue.

I reticoli degli atomi $\mathcal{R}_A \in \mathbb{R}^{\mathcal{A}_{nA}}$ e $\mathcal{R}_B \in \mathbb{R}^{\mathcal{A}_{nB}}$ sono spazi metrici con la metrica euclidea sugli atomi:

$$d_A(\alpha, \alpha') = |\mathcal{R}_A(\alpha) - \mathcal{R}_A(\alpha')| \text{ per } \alpha, \alpha' \in \mathcal{A}_{nA} \quad d_B(\beta, \beta') = |\mathcal{R}_B(\beta) - \mathcal{R}_B(\beta')| \text{ per } \beta, \beta' \in \mathcal{A}_{nB}$$

La distanza di Gromov-Hausdorff tra le firme è:

$$d_{GH}(\Sigma_A, \Sigma_B) = \inf_{\varphi: \mathcal{A}_{nA} \rightarrow \mathcal{A}_{nB}, \text{ preserva-ordine}} \sup_{\alpha \in \mathcal{A}_{nA}} |\mathcal{R}_A(\alpha) - \mathcal{R}_B(\varphi(\alpha))|$$

dove l'infimo è preso su tutti gli omomorfismi preservanti l'ordine parziale tra i reticoli delle anti-chain. Se non esistono tali omomorfismi (i reticoli hanno strutture incompatibili), la distanza è fissata a $D_{\max} = \max(\sup_{\alpha, \alpha'} |\mathcal{R}_A(\alpha) - \mathcal{R}_A(\alpha')|, \sup_{\beta, \beta'} |\mathcal{R}_B(\beta) - \mathcal{R}_B(\beta')|)$.

La distanza così definita soddisfa le proprietà di una pseudo-distanza: è simmetrica, è zero se le firme sono identiche, e soddisfa la disuguaglianza triangolare. Non è una vera distanza metrica perché due firme distinte

possono avere distanza zero se le loro strutture reticolari sono isomorfe con gli stessi pesi.

A.6 — Grado di oggettività del contenuto

Per una famiglia di N sistemi cognitivi $\{A_1, \dots, A_N\}$ e un contenuto Θ , il grado di oggettività del contenuto Θ rispetto alla famiglia è:

$$\Omega(\Theta) = 1 - (1 / N(N-1)) \cdot \sum_{i \neq j} d_{GH}(\Sigma_{\{A_i\}}(\Theta), \Sigma_{\{A_j\}}(\Theta)) / \max_{i,j} d_{GH}(\Sigma_{\{A_i\}}(\Theta), \Sigma_{\{A_j\}}(\Theta))$$

Questa quantità è compresa tra 0 e 1. Vale 1 quando tutte le firme sono identiche (massima oggettività: tutti i sistemi rappresentano Θ nella stessa struttura). Vale 0 quando le firme sono massimalmente disperse (minima oggettività: i sistemi hanno rappresentazioni di Θ strutturalmente incompatibili). Il denominatore normalizza per la massima distanza osservata nella famiglia, in modo che Ω sia sempre ben definito.

Una proprietà interessante di Ω è la dipendenza dalla famiglia: lo stesso contenuto Θ può avere Ω diversi rispetto a famiglie di sistemi diverse. Questo riflette il fatto che l'oggettività è relativa alla comunità di riferimento, non assoluta. Per contenuti scientificamente consolidati e comunicati in una comunità di esperti, Ω tende a essere alto; per contenuti con forte variabilità culturale o individuale, Ω tende a essere basso.

A.7 — Misura di errore rappresentazionale

La firma di riferimento per il contenuto Θ rispetto al sistema cognitivo A è:

$$\Sigma^*(\Theta) \equiv \operatorname{argmax}_{\{\Sigma \in \mathcal{F}(\Theta)\}} \kappa(\Sigma)$$

dove $\mathcal{F}(\Theta) = \{\Sigma(\Xi(t), \Theta) : t \in T, \text{input}(t) \text{ è istanza di } \Theta\}$ è l'insieme di tutte le firme prodotte dal sistema in risposta a input del tipo Θ nel corso del suo funzionamento.

La misura di errore rappresentazionale dello stato attuale Ξ rispetto al contenuto Θ è:

$$\text{Err}(\Xi, \Theta) = d_{GH}(\Sigma(\Xi, \Theta), \Sigma^*(\Theta))$$

Questa misura è zero quando il sistema è nella sua configurazione di massima coerenza per Θ , e cresce monotonicamente con la discrepanza tra la struttura attuale e la struttura di riferimento.

Si può mostrare che la misura di errore soddisfa una disuguaglianza triangolare modificata:

$$\text{Err}(\Xi, \Theta_1) \leq \text{Err}(\Xi, \Theta_2) + d_{\text{GH}}(\Sigma^*(\Theta_2), \Sigma^*(\Theta_1))$$

che implica che se due contenuti Θ_1 e Θ_2 hanno firme di riferimento vicine, allora l'errore relativo a Θ_1 è vicino all'errore relativo a Θ_2 . Questa proprietà è utile per estendere le previsioni da contenuti ben conosciuti a contenuti vicini ma meno familiari.

A.8 — Ricchezza rappresentazionale

La ricchezza rappresentazionale del sistema cognitivo A per il contenuto Θ è:

$$\varrho(\Xi) = \max_{\Sigma \in \mathcal{F}(\Theta)} [\kappa(\Sigma) \cdot (1 + \sigma(\Sigma))]$$

Il fattore $(1 + \sigma(\Sigma))$ amplifica la coerenza per i sistemi con alta sinergia: un sistema con κ moderata e σ alta ha ricchezza rappresentazionale maggiore di un sistema con κ alta e σ nulla, riflettendo il fatto che la sinergia aggiunge valore informativo che va oltre la semplice ridondanza.

La ricchezza rappresentazionale è monotonicamente crescente con l'apprendimento: se il sistema acquisisce nuove rappresentazioni ridondanti o approfondisce le connessioni tra quelle esistenti, il massimo su $\mathcal{F}(\Theta)$ può solo aumentare o rimanere uguale. Questo la rende una misura del progresso cognitivo per il contenuto Θ .

A.9 — Funzione di energia del sistema rappresentazionale e dinamica degli attrattori

Il sistema di rappresentazioni ridondanti $\Xi = (\xi_1, \dots, \xi_n)$ evolve secondo la dinamica:

$$d\xi_i/dt = -\partial V/\partial \xi_i + \eta_i(t)$$

dove $\eta_i(t)$ è rumore gaussiano bianco con intensità D_i , e la funzione di energia $V(\Xi)$ è:

$$V(\Xi) = -TC(\Xi \mid \Theta) + \lambda \cdot \sum_i \|\xi_i\|^2 - E\|^2$$

con $TC(\Xi \mid \Theta)$ la co-informazione totale condizionale e $\lambda > 0$ il parametro di bilanciamento tra coerenza interna e fedeltà all'input E .

I minimi locali di V sono gli attrattori del sistema. Per il sistema di Langevin con funzione di energia V , la distribuzione stazionaria è:

$$p_{\text{stat}}(\Xi) \propto \exp(-V(\Xi) / D)$$

dove $D = \sum_i D_i / n$ è il livello medio di rumore. Il sistema tende verso i minimi di V con probabilità inversa alle barriere energetiche tra i bacini.

Il tempo di transizione tra due minimi V_1 e V_2 separati da una barriera di altezza ΔV è dato approssimativamente dalla formula di Kramers:

$$\tau \approx \tau_0 \cdot \exp(\Delta V / D)$$

dove τ_0 è la scala temporale intrinseca dell'oscillazione nel bacino. Questa formula mostra che le transizioni tra stati di coerenza diversa sono esponenzialmente più lente all'aumentare delle barriere e al diminuire del rumore, il che corrisponde alla stabilità delle rappresentazioni consolidate.

A.10 — Misura di incoerenza ibrida nel cambiamento concettuale

Durante il cambiamento concettuale, la firma del sistema ha struttura ibrida: coesistono sotto-reticoli corrispondenti al vecchio framework concettuale $\mathcal{R}_{\text{old}}(\Sigma)$ e al nuovo $\mathcal{R}_{\text{new}}(\Sigma)$. La misura di incoerenza ibrida è:

$$\mathcal{H}(\Sigma) = d_{\text{GH}}(\mathcal{R}_{\text{old}}(\Sigma), \mathcal{R}_{\text{new}}(\Sigma))$$

dove d_{GH} è la distanza di Gromov-Hausdorff tra i due sotto-reticoli considerati come spazi metrici indipendenti. Questa misura è elevata quando i due sotto-reticoli hanno strutture incompatibili — quando predicono distribuzioni diverse sullo stesso spazio di stimoli — e si riduce quando uno dei due diventa dominante.

La firma ibrida può essere vista come il risultato di una sovrapposizione parziale di due attrattori della funzione di energia V : il sistema si trova in uno stato metastabile tra i due minimi, con componenti che appartengono a ciascuno. Il cambiamento concettuale completo corrisponde alla transizione da questo stato metastabile a uno dei due minimi, con dissoluzione dello stato ibrido.

A.11 — Continuità identitaria concettuale

La continuità identitaria di un concetto lungo la traiettoria $\{\Sigma(t) : t \in [t_0, t_1]\}$ nello spazio delle firme è:

$$C([\Sigma_{t_0}, \Sigma_{t_1}]) = 1 - \sup_{t \in [t_0, t_1]} \text{diam}(\mathcal{R}(\Sigma_t) \triangle \mathcal{R}(\Sigma_{t_0})) / D_{\max}$$

dove $\triangle$ è la differenza simmetrica tra i grafi orientati dei reticoli degli atomi (considerati come insiemi di archi), diam è il diametro massimo nella metrica di editing dei grafi (il numero minimo di operazioni elementari di aggiunta/rimozione di nodi e archi per trasformare un grafo nell'altro), e $D_{\max}$ è il diametro massimo possibile nello spazio dei reticoli.

La differenza simmetrica $\mathcal{R}(\Sigma_t) \triangle \mathcal{R}(\Sigma_{t_0})$ misura quali atomi informativi sono presenti in uno dei due reticoli ma non nell'altro: gli atomi che sono scomparsi dall'inizio al tempo t , più gli atomi che sono apparsi. Il diametro di questo insieme di discrepanze è la misura della lontananza strutturale tra le due firme.

Quando $\text{diam}(\mathcal{R}(\Sigma_t) \triangle \mathcal{R}(\Sigma_{t_0}))$ è piccolo per tutto $t \in [t_0, t_1]$, la continuità identitaria è alta: il concetto al tempo t_1 è ottenuto per trasformazione graduale del concetto al tempo t_0 , con conservazione della struttura reticolare essenziale. Quando diam è grande per qualche t nel periodo, la continuità è bassa: la traiettoria ha attraversato uno stato di discontinuità strutturale.

A.12 — Misura di allineamento dei sistemi artificiali

La misura globale di disallineamento di un sistema artificiale rispetto a una famiglia di sistemi biologici di riferimento per l'insieme Θ_{norm} dei concetti normativamente rilevanti è:

$$\Lambda = \Sigma_{\{\theta \in \Theta_{\text{norm}}\}} w_{\theta} \cdot d_{\text{GH}}(\Sigma_{\text{AI}}(\theta), \Sigma_{\text{human}}(\theta))$$

dove $w_{\theta} \in [0,1]$ con $\Sigma_{\{\theta\}} w_{\theta} = 1$ sono pesi che riflettono la rilevanza relativa di ciascun concetto per il dominio applicativo, $\Sigma_{\text{AI}}(\theta)$ è la firma informazionale del sistema artificiale per il concetto θ , e $\Sigma_{\text{human}}(\theta)$ è la firma media dei sistemi biologici di riferimento:

$$\Sigma_{\text{human}}(\theta) = \text{argmin}_{\{\Sigma\}} \Sigma_i d_{\text{GH}}(\Sigma, \Sigma_{\{h_i\}}(\theta))$$

dove la media è calcolata nel senso della media di Fréchet nello spazio delle firme dotato della metrica d_{GH} . Un sistema con Λ piccolo è allineato nel senso che le sue firme per i concetti normativamente rilevanti sono strutturalmente vicine alle firme umane medie; un sistema con Λ grande è disallineato anche se le sue prestazioni su benchmark standard sono eccellenti, perché la distanza strutturale delle firme produce divergenza comportamentale in condizioni non previste dall'addestramento.

A.13 — Distanza di ancoraggio causale

La distanza di ancoraggio causale misura la discrepanza tra la struttura di indipendenza condizionale della firma effettiva di un sistema e la struttura che si osserverebbe se le componenti fossero generate da canali causali genuinamente indipendenti:

$$\Delta_{\text{causal}}(\Xi, \Theta) = \|\Sigma_{\text{indep}}(\Theta) - \Sigma_{\text{corr}}(\Xi, \Theta)\|$$

dove $\Sigma_{\text{indep}}(\Theta)$ è la firma ipotetica di un sistema in cui le componenti $\xi_1, \dots, \xi_n$ sono condizionalmente indipendenti dato Θ — cioè un sistema in cui $P(\xi_1, \dots, \xi_n \mid \Theta) = \prod_i P(\xi_i \mid \Theta)$ — e $\Sigma_{\text{corr}}(\Xi, \Theta)$ è la firma effettiva del sistema che include le correlazioni residue tra le componenti dopo la condizionalizzazione su Θ .

La differenza $\Sigma_{\text{indep}} - \Sigma_{\text{corr}}$ cattura le correlazioni "spurie" tra le componenti della firma che non derivano dall'oggetto Θ nel mondo ma dalla mediazione comune del processo di apprendimento. Nei sistemi biologici, queste correlazioni spurie sono ridotte dalla varietà e dall'indipendenza fisica delle esperienze sensoriali. Nei sistemi artificiali addestrati su corpus, queste correlazioni spurie possono essere elevate perché tutte le componenti della firma derivano dalla stessa fonte.

La norma $\|\cdot\|$ è la norma euclidea sullo spazio dei vettori di atomi, opportunamente normalizzata per la dimensione del reticolo:

$$\|\Sigma_A - \Sigma_B\| = \sqrt{(1/|\mathcal{A}|) \cdot \sum_{\alpha \in \mathcal{A}} (\mathcal{R}_A(\alpha) - \mathcal{R}_B(\alpha))^2}$$

A.14 — Ricapitolazione delle relazioni tra le misure principali

Le misure principali introdotte nel quadro formale sono mutuamente coerenti nel senso seguente. La firma informazionale $\Sigma(\Xi, \Theta)$ è l'oggetto fondamentale, dal quale derivano tutte le altre misure:

La misura di errore $\text{Err}(\Xi, \Theta)$ è la distanza della firma attuale dalla firma di riferimento $\Sigma^*(\Theta)$. Quando Err è zero, il sistema è nella configurazione di massima coerenza.

La ricchezza rappresentazionale $\varrho(\Xi)$ è il valore massimo del prodotto $\kappa \cdot (1 + \sigma)$ sull'insieme delle firme producibili. Misura il potenziale rappresentazionale del sistema, non il suo stato attuale.

Il grado di oggettività $\Omega(\Theta)$ misura la dispersione delle firme tra i sistemi della famiglia di riferimento. È alto quando i sistemi convergono verso firme simili, basso quando divergono.

La continuità identitaria C misura la stabilità della struttura reticolare lungo la traiettoria temporale. Distingue il cambiamento graduale (C alta) dalla rottura concettuale (C bassa).

La distanza di ancoraggio causale Δ_{causal} distingue la firma genuinamente ancorata al mondo (Δ_{causal} piccolo) dalla firma simulata (Δ_{causal} grande). È il parametro che distingue le rappresentazioni biologiche da quelle artificiali addestrate su corpus.

La misura di disallineamento Λ aggrega le distanze delle firme artificiali dalle firme umane per i concetti normativamente rilevanti, con pesi che riflettono la rilevanza relativa di ciascun concetto.

Queste relazioni possono essere schematizzate come un sistema di dipendenze: $\Sigma \rightarrow \{\text{Err}, \kappa, \sigma, \varrho\} \rightarrow \{\Omega, C, \Delta_{\text{causal}}\} \rightarrow \Lambda$. La firma è l'oggetto primitivo; le altre misure sono funzioni della firma o di insiemi di firme; la misura di disallineamento è la più aggregata e dipende da tutte le precedenti.

A.15 — Omologia simpliciale e numeri di Betti del reticolo degli atomi

A.15,1 — Il complesso di Vietoris-Rips come filtrazione del reticolo

Sia $\mathcal{R} = (\mathcal{A}, d)$ il reticolo degli atomi PID della firma $\Sigma(\Xi, \Theta)$, dotato della metrica $d(\alpha, \beta) = d_{\text{JS}}(P_{\alpha}(\Theta), P_{\beta}(\Theta))$ tra distribuzioni condizionali degli atomi (sezione V.6,2 del testo). Il complesso di Vietoris-Rips a scala ε è:

$$\text{VR}(\mathcal{A}, \varepsilon) = \{ \sigma \subseteq \mathcal{A} : d(\alpha, \beta) \leq \varepsilon \text{ per ogni } \alpha, \beta \in \sigma \}$$

La filtrazione è la famiglia annidata:

$$\text{VR}(\varepsilon_0) \subseteq \text{VR}(\varepsilon_1) \subseteq \dots \subseteq \text{VR}(\infty) = \Delta(\mathcal{A})$$

per $\varepsilon_0 < \varepsilon_1 < \dots$ I complessi $VR(\varepsilon_i)$ crescono in modo monotono al crescere di ε_i .

Il complesso di Čech alternativo $\check{C}ech(\mathcal{A}, \varepsilon)$ soddisfa la relazione di inclusione:

$$\check{C}ech(\mathcal{A}, \varepsilon) \subseteq VR(\mathcal{A}, \varepsilon) \subseteq \check{C}ech(\mathcal{A}, 2\varepsilon)$$

per qualsiasi ε , il che garantisce che le strutture topologiche catturate dai due complessi non differiscano per più di un fattore di scala 2. Per gli scopi computazionali si usa VR; per le garanzie teoriche di omotopia si usa Čech.

A.15,2 — Operatori bordo e gruppi di omologia

Per il complesso $VR(\mathcal{A}, \varepsilon)$, siano:

C_k = Z-modulo libero generato dai k-simplessi orientati di $VR(\mathcal{A}, \varepsilon)$

L'operatore bordo $\partial_k : C_k \rightarrow C_{k-1}$ è definito su ogni k-simpleso orientato $\sigma = (v_0, v_1, \dots, v_k)$ come:

$$\partial_k(v_0, \dots, v_k) = \sum_{i=0}^k (-1)^i (v_0, \dots, \hat{v}_i, \dots, v_k)$$

dove il cappello indica l'omissione. La proprietà fondamentale è:

$$\partial_k \circ \partial_{k+1} = 0 \text{ per ogni } k \geq 0$$

che garantisce $\text{im}(\partial_{k+1}) \subseteq \ker(\partial_k)$ e rende ben definito il k-esimo gruppo di omologia:

$$H_k(VR(\mathcal{A}, \varepsilon); \mathbb{Z}) = \ker(\partial_k) / \text{im}(\partial_{k+1})$$

Il k-esimo numero di Betti è il rango del gruppo libero in H_k :

$$\beta_k(\varepsilon) = \text{rank}_{\mathbb{Z}}(H_k(VR(\mathcal{A}, \varepsilon); \mathbb{Z}))$$

Il significato geometrico per la firma informativa è il seguente: — $\beta_0(\varepsilon)$: numero di cluster di atomi semanticamente coesi alla scala ε — $\beta_1(\varepsilon)$: numero di cicli informativi irriducibili alla scala ε — $\beta_2(\varepsilon)$: numero di cavità tridimensionali (zone di incertezza strutturata)

A.15,3 — Il diagramma di persistenza e la distanza di bottleneck

L'omologia persistente traccia la nascita e la morte dei generatori di H_k al variare di ε . Il diagramma di persistenza in grado k è:

$$DP_k(\mathcal{R}) = \{ (b_i, d_i) \in \mathbb{R}^2 : b_i < d_i, \text{ generatore } i \text{ nasce a } \varepsilon = b_i \text{ e muore a } \varepsilon = d_i \}$$

La persistenza del generatore i è $\text{pers}_i = d_i - b_i$. Generatori con $\text{pers}_i > \theta$ sono considerati strutture topologiche robuste; quelli con $\text{pers}_i \leq \theta$ sono interpretati come rumore (con θ scelto mediante analisi della curva di persistenza).

La distanza di bottleneck tra diagrammi di persistenza D e D' è:

$$d_B(D, D') = \inf_{\gamma} \sup_{x \in D} \|x - \gamma(x)\|_{\infty}$$

dove l'infimo è su tutte le biiezioni $\gamma: D \cup \text{Diag} \rightarrow D' \cup \text{Diag}$ tra i diagrammi aumentati con la diagonale $\{(b, d) : b=d\}$. Ogni generatore può essere abbinato a un punto sulla diagonale (il suo "annullamento"), il che rende il confronto robusto rispetto all'aggiunta o rimozione di generatori a bassa persistenza.

Teorema di stabilità (Edelsbrunner-Harer 2008, Chazal et al. 2012): per qualsiasi coppia di spazi metrici compatti X, Y ,

$$d_B(DP_k(X), DP_k(Y)) \leq d_{GH}(X, Y) \text{ per ogni } k \geq 0$$

Applicato alle firme informative: se $d_{GH}(\Sigma, \Sigma') \leq \delta$ allora $d_B(DP_k(\Sigma), DP_k(\Sigma')) \leq \delta$. Piccole perturbazioni della struttura metrica del reticolo producono piccole variazioni nei diagrammi di persistenza.

A.15.4 — Invarianti di Betti robusti della firma

Definizione A.15.4.1 (Betti robusto di grado 0). Per soglia $\theta > 0$:

$$\beta_0^*(\Sigma, \theta) = |\{ (b_i, d_i) \in DP_0(\mathcal{R}) : d_i - b_i > \theta \}|$$

Conta i cluster di atomi che persistono per almeno θ unità di scala prima di fondersi.

Definizione A.15.4.2 (Betti robusto di grado 1). Per soglia $\theta > 0$:

$$\beta_1^*(\Sigma, \theta) = |\{ (b_i, d_i) \in DP_1(\mathcal{R}) : d_i - b_i > \theta \}|$$

Conta i cicli informativi che persistono per almeno θ unità prima di essere riempiti.

Definizione A.15.4.3 (Norma del diagramma di persistenza in grado k). Per $p \geq 1$:

$$\|DP_k(\Sigma)\|_p = (\sum_i (\text{pers}_i)^p)^{1/p}$$

Per $p = 2$, è la norma euclidea delle persistenze. Per $p = \infty$, è la persistenza massima $\max_i \text{pers}_i$.

Proprietà. Le tre funzioni β_0^* , β_1^* , $\|DP_k\|_p$ sono stabili nel senso della distanza di bottleneck: se $d_{GH}(\Sigma, \Sigma') \leq \delta$, allora $|\beta_0^*(\Sigma, \theta) - \beta_0^*(\Sigma', \theta)| \leq \delta$.

$\theta) = 0$ per $\theta > 2\delta$ (nessun generatore con persistenza $> 2\delta$ può nascere o morire in una perturbazione δ), e $|\|DP_k(\Sigma)\|_p - \|DP_k(\Sigma')\|_p| \leq C_p \cdot \delta$ per una costante C_p che dipende da p e dal numero di generatori.

A.16 — O-information e misure surrogate della firma in alta dimensionalità

A.16,1 — La O-information come misura scalare del bilanciamento ridondanza-sinergia

Per n variabili $X_1, \dots, X_n$ rispetto a un obiettivo Y , la O-information è definita come:

$$\Omega(X_1, \dots, X_n; Y) = (n-2) I(X_1, \dots, X_n; Y) + \sum_i I(X_i; Y) - \sum_i I(X_1, \dots, \hat{X}_i, \dots, X_n; Y)$$

dove $\hat{X}_i$ indica l'omissione della sorgente i . Questa quantità può essere riscritta usando l'entropia:

$$\Omega = (n-2) I(S; Y) + \sum_i I(X_i; Y) - \sum_i I(S \setminus \{X_i\}; Y)$$

dove $S = \{X_1, \dots, X_n\}$. Il segno di Ω discrimina tra sistemi prevalentemente ridondanti ($\Omega > 0$) e sistemi prevalentemente sinergici ($\Omega < 0$).

La relazione tra O-information e atomi PID è:

$$\Omega(X_1, \dots, X_n; Y) = \sum_{\alpha \in \mathcal{A}} c(\alpha) \cdot \mathcal{R}(\alpha; Y)$$

dove i coefficienti $c(\alpha) \in \{-1, 0, +1\}$ dipendono dalla struttura dell'antichain α : gli atomi di ridondanza pura hanno $c(\alpha) > 0$, gli atomi di sinergia pura hanno $c(\alpha) < 0$, gli atomi misti hanno $c(\alpha) = 0$. La O-information è dunque una proiezione scalare del vettore completo degli atomi — cattura il bilanciamento ma non la struttura dettagliata della firma.

Complessità computazionale. Il calcolo di Ω richiede $n+1$ valutazioni di informazione mutua ($I(S; Y)$ e le n informazioni mutue marginali $I(S \setminus \{X_i\}; Y)$), ciascuna delle quali richiede la stima di un'entropia congiunta di dimensione al più n . Per $n \leq 30$ e campioni di dimensione $N \geq 500$, il calcolo è fattibile con stimatori KSG in tempo $O(n^2 \cdot N \log N)$.

A.16,2 — Sinergia di coppia e approssimazione gerarchica

La sinergia di coppia $\tilde{S}(X_i, X; Y)$ per ogni coppia (i, j) è:

$$\tilde{S}(X_i, X; Y) = I(X_i, X; Y) - \max(I(X_i; Y), I(X; Y))$$

Questa quantità è sempre non-negativa ed è zero quando una delle due sorgenti cattura completamente l'informazione della coppia. La matrice di sinergia di coppia $\tilde{S} = [\tilde{S}(X_i, X; Y)]_{i,j}$ è una caratterizzazione parziale della struttura sinergica della firma, accessibile in $O(n^2)$ calcoli.

L'approssimazione gerarchica della sinergia totale è:

$$\tilde{S}_{\text{total}} \approx \sum_{i < j} \tilde{S}(X_i, X; Y)$$

che sottostima la sinergia totale perché ignora le sinergie di ordine superiore. La relazione esatta è:

$$S_{\text{total}} = \sum_{k=2}^n \sum_{|A|=k} S_k(A; Y) \geq \sum_{i < j} \tilde{S}(X_i, X; Y)$$

con uguaglianza se e solo se le sinergie di ordine $k > 2$ sono tutte zero.

A.16,3 — Lo stimatore KSG per l'informazione mutua in serie temporali continue

Per variabili aleatorie continue con campione $\{(x_i, y_i)\}_{i=1}^N$, lo stimatore di Kraskov-Stögbauer-Grassberger dell'informazione mutua $I(X; Y)$ è:

$$\hat{I}_{\text{KSG}}(X; Y) = \psi(k) + \psi(N) - \langle \psi(n_{x(i)} + 1) + \psi(n_{y(i)} + 1) \rangle_i$$

dove ψ è la funzione digamma $\psi(z) = \Gamma'(z)/\Gamma(z)$, k è il numero di vicini considerati per ciascun punto (tipicamente $k = 3, 5$, o 10), $n_{x(i)}$ è il numero di punti $j \neq i$ con $|x_j - x_i| < \varepsilon_{x(i)}$, $n_{y(i)}$ è il numero di punti $j \neq i$ con $|y_j - y_i| < \varepsilon_{y(i)}$, e $\varepsilon_{x(i)}$, $\varepsilon_{y(i)}$ sono le proiezioni sulle coordinate X, Y della distanza dal k -esimo vicino più prossimo del punto i nello spazio congiunto (X, Y) con metrica di Chebyshev.

Proprietà dello stimatore KSG. — Bias: $E[\hat{I}_{\text{KSG}}] = I(X; Y) + O(k/N)$ per densità regolari (la correzione digamma compensa il bias del primo ordine). — Varianza: $\text{Var}[\hat{I}_{\text{KSG}}] = O(1/N)$ per $N \rightarrow \infty$, con costante che dipende da k e dalla regolarità della densità. — Consistenza: $\hat{I}_{\text{KSG}} \rightarrow I(X; Y)$ in probabilità per $N \rightarrow \infty$ se la densità è lipschitziana con costante finita. — Scelta di k : $k=5$ è un valore di default robusto per N tra 200 e 10^4 ; per $N > 10^4$ è preferibile $k = \lceil \sqrt{N/10} \rceil$.

Per la stima dell'informazione mutua congiunta $I(X_1, \dots, X; Y)$, si applica lo stesso stimatore nello spazio congiunto $(X_1, \dots, X, Y)$ di dimensione $n+d_Y$ con la metrica di Chebyshev estesa a tutte le coordinate.

A.17 — La firma di riferimento deflazionista: media di Fréchet nello spazio delle firme

A.17,1 — Definizione della media di Fréchet

Sia $\{\Sigma_1, \dots, \Sigma_N\}$ una famiglia di N firme informazionali nello spazio metrico $(\mathcal{F}, d_{GH})$. La media di Fréchet della famiglia è:

$$\Sigma^* = \operatorname{argmin}_{\Sigma \in \mathcal{F}} \sum_{i=1}^N d_{GH}(\Sigma, \Sigma_i)^2$$

Questa è la firma che minimizza la somma delle distanze al quadrato dalle firme della famiglia — l'analogo della media aritmetica per spazi metrici non euclidei. Per spazi metrici curvi (come lo spazio delle firme con la metrica d_{GH}), la media di Fréchet può non essere unica e dipende dalla curvatura dello spazio.

Esistenza. In spazi metrici compatti, la media di Fréchet esiste sempre. Lo spazio delle firme $(\mathcal{F}, d_{GH})$ non è compatto in generale, ma per firme definite su un insieme finito di atomi con pesi limitati, $\mathcal{F}$ è localmente compatto e la media di Fréchet esiste sotto condizioni di crescita moderata.

Unicità. La media di Fréchet è unica in spazi con curvatura non positiva nel senso di Alexandrov (spazi $CAT(0)$). Non è noto se $(\mathcal{F}, d_{GH})$ sia uno spazio $CAT(0)$, ma risultati recenti sulla geometria degli spazi di misura suggeriscono che condizioni sufficienti per l'unicità sono soddisfatte per famiglie di firme con strutture reticolari «vicine».

A.17,2 — La firma di riferimento deflazionista per una popolazione

Dato un insieme di firme $\{\Sigma_i(\Theta)\}_{i=1}^N$ rispetto all'obiettivo Θ (una per ciascun soggetto nella popolazione di riferimento Pop), la firma di riferimento deflazionista è la media di Fréchet:

$$\Sigma^*(\Theta, \text{Pop}) = \operatorname{argmin}_{\Sigma \in \mathcal{F}} \sum_i d_{GH}(\Sigma, \Sigma_i(\Theta))^2$$

L'errore rappresentazionale dello stato Ξ del soggetto s rispetto alla firma di riferimento deflazionista è:

$$\text{Err}_{\text{pop}}(\Xi, \Theta) = d_{GH}(\Sigma_s(\Theta), \Sigma^*(\Theta, \text{Pop}))$$

Proprietà della misura deflazionista. — Relatività: Err_{pop} dipende dalla scelta di Pop. Sostituire Pop con Pop' produce in generale $\Sigma^*(\Theta, \text{Pop}') \neq \Sigma^*(\Theta, \text{Pop})$ e quindi misure di errore diverse per lo stesso sistema. — Consistenza con la media aritmetica: se lo spazio delle firme è piatto (isometrico a un sottospazio di uno spazio di Hilbert), la media

di Fréchet coincide con la media coordinata degli atomi: $\Sigma(\alpha) = (1/N) \sum_i \mathcal{R}_i(\alpha; \Theta)$ per ogni α . — Stabilità: la media di Fréchet è stabile rispetto a piccole perturbazioni delle firme individuali: se ciascuna Σ_i è perturbata di al più δ nel senso di d_{GH} , la media di Fréchet si sposta di al più δ .

A.17,3 — Il candidato punto fisso della mappa F per la firma di riferimento assoluta

Come discusso nella sezione XI.3,2, si definisce la mappa circolare:

$$F : \Sigma \mapsto \Sigma^*(\Theta(\Sigma))$$

dove $\Theta(\Sigma) = \operatorname{argmax}_Y q(\Xi, Y)$ è l'obiettivo endogeno determinato dalla firma Σ , e $\Sigma^*(\Theta)$ è la firma di riferimento per quell'obiettivo. Un punto fisso $\Sigma^{**} = F(\Sigma^{**})$ soddisfa simultaneamente le due condizioni: l'obiettivo è determinato dalla firma, e la firma è la firma di riferimento per quell'obiettivo.

Condizione sufficiente per l'esistenza. Se la mappa $\Theta(\cdot) : \mathcal{F} \rightarrow \mathcal{Y}$ è lipschitziana con costante L_Θ (piccole perturbazioni della firma producono piccole perturbazioni dell'obiettivo), e la mappa $\Sigma^*(\cdot) : \mathcal{Y} \rightarrow \mathcal{F}$ è lipschitziana con costante L_Σ (piccole perturbazioni dell'obiettivo producono piccole perturbazioni della firma di riferimento), allora F è lipschitziana con costante $L_\Theta \cdot L_\Sigma$. Se $L_\Theta \cdot L_\Sigma < 1$, il Teorema di Banach-Caccioppoli garantisce l'esistenza e l'unicità del punto fisso Σ^{**} , nonché la convergenza dell'iterazione $\Sigma_{n+1} = F(\Sigma_n)$ al punto fisso.

Interpretazione. La condizione $L_\Theta \cdot L_\Sigma < 1$ è una condizione di «risposta calibrata»: piccole variazioni della firma non producono grandi variazioni nell'obiettivo, e piccole variazioni dell'obiettivo non producono grandi variazioni nella firma di riferimento. Sistemi cognitivi robusti — sistemi in cui la struttura del contenuto non cambia drasticamente per piccole perturbazioni del sistema di rappresentazioni — soddisfano questa condizione.

A.18 — Il modello GCM-firma per la generalizzazione categoriale

A.18,1 — Specificazione formale del modello

Il Generalized Context Model con firme informative (GCM-firma) è definito come segue. Sia $C_1, \dots, C_M$ un insieme di categorie, e sia $\{x_i\}_{i=1}^N$ l'insieme degli esemplari noti con categorizzazione nota. Per ogni esemplare x_i sia Σ_i la firma informazionale. Per ogni nuovo oggetto x sia $\Sigma(x)$ la sua firma.

La probabilità di classificare x come membro della categoria C è:

$$P_{\text{firma}}(C | x) = [\sum_{\{x_i \in C\}} b_i \cdot \text{sim_firma}(x, x_i)] / [\sum_{\{x_i \in C\}} b_i \cdot \text{sim_firma}(x, x_i)]$$

dove $b_i > 0$ sono i bias delle categorie (con $\sum b_i = 1$) e la funzione di similarità è:

$$\text{sim_firma}(x, x_i) = \exp(-c \cdot d_{\text{GH}}(\Sigma(x), \Sigma_i)^r)$$

con $c > 0$ parametro di sensibilità e $r \in \{1, 2\}$ che determina la metrica ($r=1$: city-block generalizzata, $r=2$: euclidea generalizzata nel senso d_{GH}).

Parametri liberi del modello: $c, r, b_1, \dots, b_M$. Con M categorie, il modello ha $M + 2$ parametri liberi.

A.18,2 — Il grado di tipicità nel GCM-firma

Il grado di tipicità dell'esemplare x_i nella categoria C è definito come:

$$\text{typ_firma}(x_i, C) = \sum_{\{x \in C, j \neq i\}} \text{sim_firma}(x_i, x) / [\sum_{\{x \in C\}} \text{sim_firma}(x_i, x)]$$

Questa è la similarità media di x_i con gli altri membri di C normalizzata per la similarità totale di x_i con tutti gli esemplari. È alta per gli esemplari centrali (simili a molti altri membri della categoria) e bassa per gli esemplari periferici (dissimili dalla maggior parte degli altri membri).

A.18,3 — La struttura sinergica residua come predittore degli errori

La struttura sinergica residua di x rispetto a C è:

$$\Delta S(x, C) = \|\text{Syn}(\Sigma(x)) - \text{Syn}(\Sigma(C))\|_2$$

dove $\text{Syn}(\Sigma) = \{S_k(\Sigma) : k = 2, \dots, n\}$ è il vettore delle componenti sinergiche di ordine k nella firma Σ , e $\text{Syn}(\Sigma(C))$ è il vettore delle componenti sinergiche della firma media della categoria. La norma $\|\cdot\|_2$ è la norma euclidea sul vettore delle sinergie.

Previsione empirica centrale. Per esemplari x che cadono vicino al confine di classificazione ($P_{\text{firma}}(C | x) \in (0,3, 0,7)$), la probabilità di errore di classificazione è una funzione crescente di $\Delta S(x, C)$:

$$P(\text{errore} | x, \Delta S) \approx P_0 + \alpha \cdot \Delta S(x, C)$$

dove P_0 è la probabilità di base per gli esemplari di frontiera senza struttura sinergica divergente, e $\alpha > 0$ è il coefficiente di influenza della struttura sinergica. Questa previsione è distinguibile da quella del GCM

standard ($\alpha = 0$) ed è falsificabile con paradigmi sperimentali che manipolano ΔS attraverso la scelta degli stimoli.

A.18,4 — Il GCM-firma con dipendenza dallo scopo

Per incorporare la dipendenza dello scopo S nella categorizzazione, la firma dell'esemplare è calcolata rispetto all'obiettivo dello scopo Y_S :

$$\Sigma_S(x) = \Sigma(\Xi(x), Y_S)$$

dove Y_S è la variabile del mondo che il classificatore intende informarsi attraverso la categorizzazione. La probabilità di classificazione dipendente dallo scopo è:

$$P_{\text{firma_scopo}}(C \mid x, S) = \left[\sum_{\{x_i \in C\}} b \cdot \exp(-c \cdot d_{\text{GH}}(\Sigma_S(x), \Sigma_S(x_i))) \right] / \left[\sum_{\{x_i \in C\}} b \cdot \exp(-c \cdot d_{\text{GH}}(\Sigma_S(x), \Sigma_S(x_i))) \right]$$

Previsione. Per due scopi $S_1 \neq S_2$ con obiettivi $Y_{\{S_1\}} \neq Y_{\{S_2\}}$, il modello predice che $P_{\text{firma_scopo}}(C \mid x, S_1) \neq P_{\text{firma_scopo}}(C \mid x, S_2)$ per (quasi) tutti gli esemplari x e categorie C . Il disaccordo di classificazione tra classificatori con scopi diversi è predetto essere una funzione crescente di $d_{\text{GH}}(\Sigma_{\{S_1\}}(x), \Sigma_{\{S_2\}}(x))$ — la distanza tra le firme dello stesso oggetto rispetto ai due diversi obiettivi.

A.19 — Proprietà formali della decomposizione PID: confronto tra le misure $I_{\min}$ e I_{BROJA}

A.19,1 — Sistema di assiomi comune

Entrambe le misure $I_{\min}$ e I_{BROJA} soddisfano il seguente sistema di assiomi base (Williams-Beer 2010):

Assioma S (Simmetria): $\mathcal{R}(\alpha; Y)$ è invariante sotto permutazioni delle etichette delle sorgenti.

Assioma M (Monotonicità): se $\alpha \leq \beta$ nel reticolo delle antichain, allora $\sum_{\{\gamma \leq \alpha\}} \mathcal{R}(\gamma; Y) \leq \sum_{\{\gamma \leq \beta\}} \mathcal{R}(\gamma; Y)$.

Assioma SR (Self-redundancy): per una singola sorgente X con antichain singleton $\{i\}$, $\mathcal{R}(\{i\}; Y) = I(X_i; Y)$.

Assioma LP (Local positivity): $\mathcal{R}(\alpha; Y) \geq 0$ per ogni α (richiesto da I_{BROJA} , violato da $I_{\min}$ per $n \geq 3$ in certi casi).

A.19,2 — Il caso XOR generalizzato come discriminatore

Il sistema XOR generalizzato a $n=3$ sorgenti: $X_1 \sim \text{Bernoulli}(0,5)$, $X_2 \sim \text{Bernoulli}(0,5)$, $X_3 = X_1 \text{ XOR } X_2$ (con $p(\text{flip}) = 0$), $Y = X_1$.

Per questo sistema: $I(X_1; Y) = 1$ bit, $I(X_2; Y) = 0$ bit, $I(X_3; Y) = 0$ bit
 $I(X_1, X_2; Y) = 1$ bit, $I(X_1, X_3; Y) = 1$ bit, $I(X_2, X_3; Y) = 1$ bit $I(X_1, X_2, X_3; Y) = 1$ bit

Con la misura $I_{\min}$: $R(\{1\} \wedge \{2\} \wedge \{3\}; Y) = \min(1, 0, 0) = 0$ $U(X_1; Y) = 1 - 0 = 1$ bit $U(X_2; Y) = 0 - 0 = 0$ bit $U(X_3; Y) = 0 - 0 = 0$ bit

La sinergia di terzo ordine richiesta per la consistenza marginale: $I(X_2, X_3; Y) = 1 = R(\{2,3\}; Y) + \dots$ e siccome $I_{\min}(\{2,3\}; Y) = I_{\min} \min(0,0) = 0$, serve una sinergia di coppia $S(\{2,3\}; Y) = 1$ bit. Ma questa sinergia è poi inconsistente con i margini di terzo ordine: produce un atomo negativo per le sinergie miste. In questo caso $I_{\min}$ viola LP.

Con la misura I_{BROJA} , si risolve il problema di ottimizzazione su $\Delta(P)$ e si ottiene: $R_{\text{BROJA}}(\{1\} \wedge \{2,3\}; Y) = I_{\min}(X_1; Y) = 1$ bit Tutti gli altri atomi di ridondanza sono zero. $U_{\text{BROJA}}(X_2; Y) = U_{\text{BROJA}}(X_3; Y) = 0$ bit. $S_{\text{BROJA}}(\{2,3\}; Y) = 1$ bit (la coppia $\{X_2, X_3\}$ porta sinergia perché $X_3 = f(X_2, Y)$).

La firma I_{BROJA} per questo sistema è dunque: alta ridondanza tra X_1 e la coppia $\{X_2, X_3\}$, alta sinergia nella coppia $\{X_2, X_3\}$, zero unicità per X_2 e X_3 .

A.19,3 — Implicazione per la firma informazionale

Le due misure producono firme diverse per lo stesso sistema. Per la teoria del contenuto, la scelta tra $I_{\min}$ e I_{BROJA} corrisponde a due diverse operativizzazioni della nozione di «ridondanza costitutiva»:

$I_{\min}$: ridondanza conservativa. Due sorgenti sono ridondanti se ciascuna porta individualmente la stessa informazione. In un sistema neurale, corrisponde alla ridondanza di tipo «backup»: due aree cerebrali che codificano indipendentemente lo stesso segnale.

I_{BROJA} : ridondanza costruttiva. Due sorgenti sono ridondanti se la struttura delle loro distribuzioni marginali con l'obiettivo è compatibile con informazione condivisa, anche se le distribuzioni di dipendenza interna tra le sorgenti sono diverse. Corrisponde alla ridondanza di tipo «coordinato»: due aree cerebrali che coordinano la loro attività per portare la stessa informazione attraverso meccanismi diversi.

Per i sistemi neurobiologici in cui le aree cerebrali sono connesse da fascicoli anatomici e la loro attività è correlata attraverso oscillazioni di fase, la ridondanza costruttiva di I_{BROJA} è teoricamente più adeguata.

Per i sistemi artificiali con canali di elaborazione progettati per essere indipendenti, la ridondanza conservativa di $I_{\min}$ è più appropriata.

A.20 — I numeri di Dedekind e la complessità combinatoria del reticolo degli atomi PID

A.20,1 — Definizione e valori

Il numero di Dedekind $D(n)$ conta il numero di antichain dell'insieme delle parti $\mathcal{P}(\{1, \dots, n\})$ — ovvero il numero di insiemi di sottoinsiemi non vuoti di $\{1, \dots, n\}$ in cui nessun elemento è sottoinsieme di un altro. I valori esatti per $n \leq 8$ sono:

$$\begin{array}{llllllll} D(0) = 2 & D(1) = 3 & D(2) = 6 & D(3) = 20 & D(4) = 168 & D(5) = 7,581 & D(6) = & \\ 7,828,354 & & D(7) = & & 2,414,682,040,998 & & D(8) = & \\ 56,130,437,228,687,557,907,788 & & & & & & & \end{array}$$

Per $n = 9$, il valore è dell'ordine di 10^{23} ; per $n = 10$, è dell'ordine di 10^{28} . La crescita è asintoticamente:

$$\log_2 D(n) \sim C(n, \lfloor n/2 \rfloor) = n! / (\lfloor n/2 \rfloor! \cdot \lceil n/2 \rceil!)$$

dove $C(n, \lfloor n/2 \rfloor)$ è il coefficiente binomiale centrale, che cresce come $2^n / \sqrt{(\pi n/2)}$.

A.20,2 — Implicazioni per la complessità computazionale della PID

Il reticolo degli atomi PID per n sorgenti ha $|\mathcal{A}|$ elementi. Il calcolo della decomposizione esatta richiede:

— per $i_{\min}$: $O(D(n))$ valutazioni di informazione puntuale $I(X_i; Y=y)$ e il calcolo del minimo su ciascuna antichain, con costo totale $O(D(n) \cdot |\mathcal{Y}| \cdot n)$ dove $\mathcal{Y}$ è il supporto dell'obiettivo. — per I_{BROJA} : $O(D(n))$ problemi di ottimizzazione convessa, ciascuno di dimensione $O(|\mathcal{X}_1| \cdot |\mathcal{X}_2| \cdot |\mathcal{Y}|)$, con costo totale molto più elevato.

Per i valori pratici di n ($n \leq 4$: $D(4) = 168$, accessibile; $n = 5$: $D(5) = 7581$, marginalmente accessibile; $n \geq 6$: computazionalmente intractable per il calcolo esatto), la PID esatta può essere calcolata solo per sistemi di piccola dimensionalità.

A.20,3 — Strategie di approssimazione gerarchica e loro precisione

L'approssimazione gerarchica a coppie calcola le PID binarie $\{\mathcal{R}(\{i\} \wedge \{j\}; Y) : i < j\}$ in $O(n^2)$ calcoli e poi riassume la struttura della firma attraverso la matrice di ridondanza binaria. La perdita di precisione rispetto alla PID esatta è quantificata dalla sinergia di ordine superiore non catturata:

$$\Delta_{\text{gerarchica}} = I(X_1, \dots, X; Y) - \max_{|A|=2} I(X_A; Y) - \sum_{\{i,j\}} S(\{i,j\}; Y)$$

che misura l'informazione non catturata da nessuna analisi a coppie. Per sistemi con $\Delta_{\text{gerarchica}} / I(X_1, \dots, X; Y) < 0,1$, l'approssimazione a coppie preserva il 90% della struttura informazionale.

A.21 — Protocollo standard di stima della firma informazionale da dati neuroimaging fMRI

A.21,1 — Pre-processing e definizione delle variabili

Passo 1 — Scelta delle ROI. Selezionare $k \geq 3$ regioni di interesse (ROI) sulla base di ipotesi anatomiche e funzionali sul sistema cognitivo in studio. Per ciascuna ROI, estrarre la serie temporale media del segnale BOLD dopo pre-processing standard (correzione del movimento, correzione della disomogeneità di campo, normalizzazione spaziale al template MNI).

Passo 2 — Definizione dell'obiettivo. Definire la variabile obiettivo Y come la variabile dello stimolo o del compito rilevante per la firma (es. identità dello stimolo, categoria semantica, valenza emotiva). Costruire la serie temporale di Y convolvendo la sequenza degli stimoli con la funzione di risposta emodinamica (HRF) canonica.

Passo 3 — Riduzione del rumore. Applicare una componente di regressione per i segnali di confound (movimento, segnale del ventricolo, segnale della white matter) per ridurre le correlazioni spurie tra le ROI non dovute all'elaborazione cognitiva.

A.21,2 — Stima delle informazioni mutue con il metodo KSG

Per ciascuna coppia (ROI_i, Y) , stimare $I(ROI_i; Y)$ con lo stimatore KSG:

$$\hat{I}(ROI_i; Y) = \psi(k) + \psi(N) - \langle \psi(n_{\{ROI_i\}}(t) + 1) + \psi(n_Y(t) + 1) \rangle_t$$

con $k = 5$ come valore di default e N = numero di volumi temporali (tipicamente $N = 200-600$ per una sessione fMRI di 10–30 minuti).

Per la stima dell'informazione mutua condizionale $I(\text{ROI}_i; Y | \text{ROI}_j)$, usare il metodo KSG nello spazio congiunto $(\text{ROI}_j, \text{ROI}_i, Y)$ o il metodo dei kernel a variazione lenta (CMI-kNN) che ha migliore performance per $N < 500$.

Controllo per dipendenze seriali. Le serie temporali fMRI hanno autocorrelazione temporale significativa (specialmente dopo smoothing spaziale). Correggere la stima di N con il numero effettivo di campioni indipendenti: $N_{\text{eff}} = N / (1 + 2 \sum_{l=1}^L \rho_l)$, dove ρ_l è l'autocorrelazione al lag l e L è scelto per includere i lag significativi.

A.21,3 — Calcolo della O-information e della sinergia di coppia

Dall'insieme delle informazioni mutue stimate, calcolare:

$$\Omega = (k-2) \hat{I}(\text{ROI}_1, \dots, \text{ROI}_k; Y) + \sum_i \hat{I}(\text{ROI}_i; Y) - \sum_i \hat{I}(\text{ROI}_1, \dots, \text{ROI}_k, \text{ROI}_i; Y)$$

$$\tilde{S}(\text{ROI}_i, \text{ROI}_j; Y) = \hat{I}(\text{ROI}_i, \text{ROI}_j; Y) - \max(\hat{I}(\text{ROI}_i; Y), \hat{I}(\text{ROI}_j; Y))$$

Test statistico per Ω . Costruire la distribuzione nulla di Ω attraverso il metodo delle permutazioni della serie temporale di Y (distruzione delle dipendenze temporali tra stimoli e risposte, conservando le strutture di autocorrelazione intra-ROI). La significatività è $\alpha = 0,05$ con correzione per confronti multipli (FDR sulla matrice di sinergia di coppia).

A.21,4 — Costruzione della firma ridotta e validazione

Firma ridotta. Raccogliere le stime in un vettore firma ridotto:

$$\tilde{\Sigma} = (\hat{I}(\text{ROI}_1; Y), \dots, \hat{I}(\text{ROI}_k; Y), \tilde{S}(\text{ROI}_i, \text{ROI}_j; Y) \text{ per } i < j, \Omega)$$

di dimensione $k + C(k, 2) + 1$.

Validazione interna. Dividere la sessione in due metà temporalmente indipendenti. Stimare $\tilde{\Sigma}_A$ e $\tilde{\Sigma}_B$ per le due metà. La correlazione di Pearson tra $\tilde{\Sigma}_A$ e $\tilde{\Sigma}_B$ (trattando le componenti della firma come vettori) deve essere $r > 0,7$ per una firma considerata affidabile.

Validazione predittiva. Stimare la capacità predittiva della firma $\tilde{\Sigma}$ rispetto al comportamento del soggetto in un compito indipendente attraverso la correlazione entre $\tilde{\Sigma}$ e il profilo di generalizzazione categoriale del soggetto. Un coefficiente di determinazione $R^2 > 0,3$ indica firma predittivamente valida.

A.22 — Dizionario delle notazioni

Per comodità si raccolgono qui tutte le notazioni utilizzate nelle appendici A.1–A.21 e nel testo principale.

Variabili e spazi di base

$X_1, \dots, X$ — sorgenti di informazione (rappresentazioni ridondanti) Y, Θ — obiettivo / contenuto rappresentato $\mathcal{X}_i$ — supporto della sorgente X_i $\mathcal{Y}, \Theta$ — supporto dell'obiettivo $P(\cdot), p(\cdot)$ — distribuzione di probabilità / densità di probabilità $\Delta(\Omega)$ — simpleso delle distribuzioni di probabilità su Ω

Misure di informazione

$H(X)$ — entropia di Shannon di X (in bit) $H(X|Y)$ — entropia condizionale di X dato Y $I(X;Y)$ — informazione mutua tra X e Y $I(X;Y|Z)$ — informazione mutua condizionale dato Z $C(X_1; \dots; X; Y)$ — co-informazione di ordine $n+1$ $JSD(p||q)$ — divergenza di Jensen-Shannon tra p e q $d_{JS}(p,q)$ — distanza di Jensen-Shannon = $\sqrt{JSD(p||q)}$ $d_{FR}(p,q)$ — distanza di Fisher-Rao tra distribuzioni $KL(p||q)$ — divergenza di Kullback-Leibler di p rispetto a q

Decomposizione PID

$\mathcal{A}$ — insieme delle antichain del reticolo degli ordini di $\{1, \dots, n\}$ α, β — antichain generiche in $\mathcal{A}$ $\mathcal{R}(\alpha; Y)$ — atomo PID corrispondente all'antichain α rispetto a Y $U_i(Y)$ — informazione unica della sorgente i rispetto a Y $\text{Syn}(\{1, \dots, n\}; Y)$ — sinergia di ordine n $I_{\min}$ — misura PID di Williams-Beer I_{BROJA} — misura PID di Bertschinger et al. (ottimizzazione convessa) $\Delta(P)$ — insieme delle distribuzioni con marginali bilaterali fissate $\Omega(X_1, \dots, X; Y)$ — O-information $\tilde{S}(X_i, X; Y)$ — sinergia di coppia tra X_i e X rispetto a Y $D(n)$ — numero di Dedekind (cardinalità di $\mathcal{A}$)

Firma informazionale

$\Xi = (\xi_1, \dots, \xi)$ — vettore di rappresentazioni ridondanti $\Sigma(\Xi, \Theta) = (\mathcal{R}(\Xi, \Theta), \kappa(\Xi, \Theta), \sigma(\Xi, \Theta))$ — firma informazionale $\mathcal{R}(\Xi, \Theta)$ — reticolo degli atomi della firma $\kappa(\Xi, \Theta)$ — coefficiente di coerenza della firma $\sigma(\Xi, \Theta)$ — rapporto sinergico della firma $\varrho(\Xi) = \kappa \cdot (1 + \sigma)$ — ricchezza rappresentazionale $\Sigma^*(\Theta)$ — firma di riferimento per il contenuto Θ $\Sigma^*(\Theta, \text{Pop})$ — firma di riferimento deflazionista per la popolazione Pop Σ — media di Fréchet di una famiglia di firme $\mathcal{F}$ — spazio delle firme informazionali

Distanze e metriche

$d_{GH}(\Sigma_A, \Sigma_B)$ — distanza di Gromov-Hausdorff tra firme $d_B(D, D')$ — distanza di bottleneck tra diagrammi di persistenza $d_H(A, B)$ —

distanza di Hausdorff tra insiemi nello stesso spazio $\Omega(\Theta)$ — grado di oggettività del contenuto Θ rispetto a una famiglia

Dinamica degli attrattori

$V(\Xi)$ — funzione di energia del sistema rappresentazionale $C(\Xi|\Theta)$ — co-informazione totale condizionale (termine di coerenza in V) ΔV — altezza della barriera tra bacini di attrazione $\tau \approx \tau_0 \cdot \exp(\Delta V/D)$ — tempo di transizione (formula di Kramers) $D = \sum_i D_i/n$ — intensità media del rumore

Errore e normatività

$\text{Err}(\Xi, \Theta) = d_{\text{GH}}(\Sigma(\Xi, \Theta), \Sigma^*(\Theta))$ — misura di errore rappresentazionale $\mathcal{H}(\Sigma) = d_{\text{GH}}(\mathcal{R}_{\text{old}}(\Sigma), \mathcal{R}_{\text{new}}(\Sigma))$ — incoerenza ibrida nel cambiamento concettuale $C([\Sigma_{\{t_0\}}, \Sigma_{\{t_1\}}])$ — continuità identitaria lungo la traiettoria $\Delta_{\text{causal}}(\Xi, \Theta)$ — distanza di ancoraggio causale Λ — misura globale di disallineamento dei sistemi artificiali

Topologia persistente

$\text{VR}(\mathcal{A}, \epsilon)$ — complesso di Vietoris-Rips a scala ϵ C_k — k -esimo modulo delle catene del complesso $\partial k : C_k \rightarrow C_{k-1}$ — operatore bordo $H_k(K; \mathbb{Z}) = \ker(\partial k) / \text{im}(\partial_{k+1})$ — k -esimo gruppo di omologia $\beta_k(\epsilon) = \text{rank } H_k(\text{VR}(\mathcal{A}, \epsilon); \mathbb{Z})$ — k -esimo numero di Betti alla scala ϵ $\text{DP}_k(\mathcal{R}) = \{(b_i, d_i)\}$ — diagramma di persistenza in grado k $\text{pers}_i = d_i - b_i$ — persistenza del generatore i -esimo $\beta_k^*(\Sigma, \theta)$ — numero di Betti robusto alla soglia θ $\|\text{DP}_k(\Sigma)\|_p$ — norma del diagramma di persistenza

Generalizzazione categoriale

$\Sigma(C)$ — firma della categoria C (media di Fréchet degli esemplari) $F(C) = \{\Sigma(x_i) : x_i \in C\}$ — firma distribuzionale della categoria $\text{typ_firma}(x, C)$ — grado di tipicità di x in C nel GCM-firma $\Delta S(x, C)$ — struttura sinergica residua di x rispetto a C $P_{\text{firma}}(C | x)$ — probabilità di classificazione nel GCM-firma c — parametro di sensibilità del GCM r — esponente della metrica nel GCM ($r=1$ o $r=2$)

Neuroimaging

ROI — regione di interesse cerebrale BOLD — segnale fMRI Blood-Oxygen-Level-Dependent $\hat{I}_{\text{KSG}}$ — stimatore KSG dell'informazione mutua N_{eff} — numero effettivo di campioni indipendenti $\tilde{\Sigma}$ — firma ridotta da neuroimaging $\psi(z)$ — funzione digamma

Bibliografia critica e ragionata

Le voci bibliografiche sono raggruppate tematicamente, con una breve nota che specifica il ruolo di ciascuna nell'economia dell'argomentazione. All'interno di ciascuna sezione l'ordine è cronologico. Per le opere con edizioni tradotte si indica la traduzione italiana quando disponibile; si riportano le pagine di riferimento quando la citazione riguarda una sezione specifica.

I. Teoria dell'informazione e decomposizione dell'informazione parziale

La letteratura fondativa del quadro formale. Le voci di Williams-Beer e dei successivi sviluppi della PID sono quelle a cui si fa più diretto riferimento nel Capitolo IV e nell'Appendice matematica.

Shannon, Claude E. (1948). «A Mathematical Theory of Communication». *Bell System Technical Journal*, 27(3), 379–423; 27(4), 623–656. Trad. it. parziale in: Shannon, C. E., Weaver, W. (1971). *La teoria matematica delle comunicazioni*. Milano: Etas Kompass. — La fonte della teoria dell'entropia e dell'informazione mutua come primitivi del quadro formale (Appendice A.1). L'assiomatizzazione di Shannon dell'entropia come unica misura soddisfacente continuità, simmetria e additività è il modello a cui si aspira un'assiomatizzazione analoga della PID.

Cover, Thomas M., Thomas, Joy A. (2006). *Elements of Information Theory*. Second edition. Hoboken: Wiley-Interscience. — Il testo di riferimento standard per le proprietà dell'entropia, dell'informazione mutua condizionale, della co-informazione (Capitolo 2) e delle catene di Markov informazionali. Le disuguaglianze fondamentali citate nel testo (chain rule, data processing inequality) sono riportate nei Teoremi 2.5.2 e 2.8.1.

Williams, Paul L., Beer, Randall D. (2010). «Nonnegative Decomposition of Multivariate Information». *arXiv:1004.2515*. — L'articolo fondativo della PID. Introduce le antichain come struttura reticolare per la decomposizione, la misura $I_{\min}$, e il sistema di assiomi di simmetria, monotonicità e self-redundancy. Il problema della non-negatività per $n \geq 3$ è riconosciuto dagli stessi autori nella sezione 4.

Bertschinger, Nils, Rauh, Johannes, Olbrich, Eckehard, Jost, Jürgen, Ay, Nihat (2014). «Quantifying Unique Information». *Entropy*, 16(4), 2161–2183. — Introduce la misura I_{BROJA} attraverso il problema di ottimizzazione convessa su $\Delta(P)$. La prova di non-negatività è nel Teorema 3; la proprietà di unica soluzione per il problema convesso è nel Corollario 2. La variante I_{ccs} (informazione minima comune

sorprendente) è introdotta nella sezione 3,4 e soddisfa proprietà assiomatiche diverse.

Griffith, Virgil, Koch, Christof (2014). «Quantifying Synergistic Mutual Information». In: Prokopenko, M. (ed.), *Guided Self-Organization: Inception*. Berlin: Springer, 159–190. — Introduce la misura I_{sx} definita a partire dalla sinergia globale. Il confronto sistematico con I_{min} è nella sezione 4; l'applicazione alla neuroscienza computazionale nella sezione 5.

Rauh, Johannes, Bertschinger, Nils, Olbrich, Eckehard, Jost, Jürgen (2014). «Reconsidering Unique Information: Towards a Multivariate Information Decomposition». In: *Proceedings of the IEEE International Symposium on Information Theory (ISIT 2014)*, 2232–2236. — Fornisce la prova che l'assioma di non-negatività (LP) è incompatibile con alcuni altri assiomi naturali per $n \geq 3$, suggerendo che il problema dell'unicità della PID è strutturalmente non riducibile a scelte tecniche. Rilevante per la sezione XI.4,2 del testo.

Bertschinger, Nils, Rauh, Johannes (2014). «The Shared Information of Walsh Functions». In: *Proceedings of ISIT 2014*, 2173–2177. — Calcolo esplicito della PID per sistemi con struttura di Walsh. Fornisce casi test concreti per le diverse misure e mostra che I_{min} e I_{BROJA} divergono per sistemi con strutture di dipendenza non banali.

Rosas, Fernando E., Mediano, Pedro A. M., Gastpar, Michael, Jensen, Henrik J. (2019). «Quantifying High-order Interdependencies via Multivariate Extensions of the Mutual Information». *Physical Review E*, 100(3), 032305. — Introduce la O-information come misura scalare del bilanciamento ridondanza-sinergia (Appendice A.16,1) e ne studia le proprietà asintotiche. La dimostrazione che Ω è una combinazione lineare degli atomi PID con coefficienti ± 1 è nella Proposizione 2.

Varley, Thomas F., Sporns, Olaf, Schaffelhofer, Stefan, Scherberger, Hamidreza, Dann, Benjamin (2023). «Information-theoretic Signatures of Cortical Representation of Sensorimotor Trajectories». *Nature Communications*, 14, 4971. — Applicazione empirica della O-information a dati di registrazione multi-elettrodo nella corteccia motoria. Illustra il protocollo di stima discusso nella sezione IV.5bis del testo e la procedura di validazione through-subjects.

II. Geometria dell'informazione e metriche sullo spazio delle distribuzioni

Amari, Shun-ichi (2016). *Information Geometry and Its Applications*. Tokyo: Springer Japan. — Il trattamento sistematico della geometria riemanniana sullo spazio delle distribuzioni di probabilità, con la metrica

di Fisher come struttura fondamentale. La dimostrazione dell'invarianza della distanza di Fisher-Rao sotto trasformazioni sufficienti è nel Teorema 2,1.

Ay, Nihat, Jost, Jürgen, Lê, Hồng Vân, Schwachhöfer, Lorenz (2017). *Information Geometry*. Cham: Springer. — Presentazione moderna della geometria dell'informazione con attenzione alle varietà statistiche esponenziali e miste. Rilevante per la struttura della metrica sullo spazio delle firme informazionali (sezione V.2).

Lin, Jianhua (1991). «Divergence Measures Based on the Shannon Entropy». *IEEE Transactions on Information Theory*, 37(1), 145–151. — Introduce la divergenza di Jensen-Shannon e ne prova la simmetria e la limitatezza superiore. La prova che $\sqrt{JSD}$ è una vera distanza metrica (non solo una pseudo-distanza) è nella Proposizione 3.

Endres, Dominik M., Schindelin, Johannes E. (2003). «A New Metric for Probability Distributions». *IEEE Transactions on Information Theory*, 49(7), 1858–1860. — Prova rigorosa che la radice quadrata della divergenza di Jensen-Shannon soddisfa la disuguaglianza triangolare e costituisce quindi una distanza metrica. Rilevante per la costruzione della metrica sul reticolo degli atomi nella sezione V.6,2.

III. Topologia algebrica computazionale e omologia persistente

Edelsbrunner, Herbert, Letscher, David, Zomorodian, Afra (2002). «Topological Persistence and Simplification». *Discrete & Computational Geometry*, 28(4), 511–533. — L'articolo fondativo dell'omologia persistente. Introduce la filtrazione di Vietoris-Rips, i diagrammi di persistenza, e la prova di stabilità per perturbazioni della funzione di filtrazione. La costruzione degli operatori bordo è nel paragrafo 2.

Zomorodian, Afra, Carlsson, Gunnar (2005). «Computing Persistent Homology». *Discrete & Computational Geometry*, 33(2), 249–274. — Algoritmo efficiente per il calcolo dei diagrammi di persistenza. La complessità computazionale dell'algoritmo è $O(m^3)$ dove m è il numero di semplici, il che è fattibile per complessi di dimensione moderata.

Cohen-Steiner, David, Edelsbrunner, Herbert, Harer, John (2007). «Stability of Persistence Diagrams». *Discrete & Computational Geometry*, 37(1), 103–120. — Prova originale del Teorema di Stabilità per l'omologia persistente. La versione per la distanza di Gromov-Hausdorff (citata come applicazione alla sezione V.6,3) è la Proposizione 3,3.

Chazal, Frédéric, Cohen-Steiner, David, Glisse, Marc, Guibas, Leonidas J., Oudot, Steve Y. (2009). «Proximity of Persistence Modules and Their

Diagrams». In: *Proceedings of the 25th Annual Symposium on Computational Geometry*, 237–246. — Estensione del Teorema di Stabilità ai moduli di persistenza astratti, con applicazione alla distanza di bottleneck per complessi di Vietoris-Rips su spazi metrici.

Carlsson, Gunnar (2009). «Topology and Data». *Bulletin of the American Mathematical Society*, 46(2), 255–308. — Rassegna accessibile dei metodi della TDA (Topological Data Analysis) con applicazioni in biologia, neuroimaging e analisi dei dati ad alta dimensionalità. Rilevante come contesto per l'applicazione della topologia persistente alle firme informazionali nella sezione V.6.

Edelsbrunner, Herbert, Harer, John L. (2010). *Computational Topology: An Introduction*. Providence: American Mathematical Society. — Trattazione completa della topologia computazionale inclusi i fondamenti dell'omologia simpliciale (Capitoli 3–4), l'omologia persistente (Capitolo 7) e la distanza di bottleneck (Capitolo 8). Testo di riferimento per le dimostrazioni delle proprietà degli invarianti topologici nella sezione A.15.

Oudot, Steve Y. (2015). *Persistence Theory: From Quiver Representations to Data Analysis*. Providence: AMS. — Trattazione algebrica approfondita della teoria dei moduli di persistenza, con discussione dell'omologia a parametri multipli (multi-parameter persistence) rilevante per la sezione V.6,12.

IV. Filosofia del contenuto mentale: teorie standard

Brentano, Franz (1874). *Psychologie vom empirischen Standpunkt*. Leipzig: Duncker & Humblot. Trad. it.: (1997). *La psicologia dal punto di vista empirico*. Roma-Bari: Laterza. — La fonte storica della nozione di intenzionalità come struttura fondamentale della vita mentale. Il concetto di «inesistenza intenzionale» dell'oggetto del pensiero è nell'Introduzione e nel Libro II, Capitolo 1.

Husserl, Edmund (1900–1901). *Logische Untersuchungen*. Halle: Niemeyer. Trad. it.: (1968). *Ricerche logiche*, 2 voll. Milano: Il Saggiatore. — L'analisi delle strutture intenzionali attraverso la distinzione atto/contenuto/oggetto. La nozione di «orizzonte» dell'oggetto intenzionale è introdotta nella VI Ricerca; la struttura della sintesi passiva è sviluppata nelle *Analysen zur passiven Synthesis* (Husserliana XI). Rilevante per la sezione IX.4.

Fodor, Jerry A. (1975). *The Language of Thought*. New York: Crowell. Trad. it.: (1988). *Il linguaggio del pensiero*. Milano: Guerini. — L'introduzione del

LOT (Language of Thought) come ipotesi sul substrato cognitivo del contenuto proposizionale. La struttura combinatoria dei simboli mentali è discussa nel Capitolo 2; la relazione con la modularità nella Prefazione.

Fodor, Jerry A. (1983). *The Modularity of Mind*. Cambridge, MA: MIT Press. Trad. it.: (1988). *La mente modulare*. Bologna: Il Mulino. — La tesi dell'incapsulamento informazionale dei sistemi modulari. I criteri di modularità sono nel Capitolo 3; la distinzione moduli/sistemi centrali nel Capitolo 4. La difficoltà del «problema dei sistemi centrali» per la composizionalità è discussa nel Capitolo 5 ed è rilevante per la sezione III.4 del testo.

Fodor, Jerry A. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press. — La formulazione sistematica del realismo intenzionale e del problema della proiettabilità. La critica all'olismo semantico e la difesa dell'atomismo concettuale sono nel Capitolo 3.

Dretske, Fred (1981). *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press. — L'introduzione del framework dretskiano dell'informazione come base semantica. La nozione di «sistema segnalante» e la relazione nomologica tra segnale e contenuto sono nel Capitolo 3; il problema del disjunction è affrontato nella sezione 3,1.

Dretske, Fred (1988). *Explaining Behavior: Reasons in a World of Causes*. Cambridge, MA: MIT Press. Trad. it.: (1994). *Naturalizzare la mente*. Torino: Rosenberg & Sellier. — La maturazione della teoria dretskiana con la distinzione tra informazione indicatore e funzione segnalante acquisita nell'apprendimento. Il «problema della molteplice indicazione» è discusso nel Capitolo 4. Rilevante per la sezione III.3.

Millikan, Ruth G. (1984). *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press. — La fonte della teleosemantica. La nozione di «funzione propria» è definita nel Capitolo 1; il sistema produttore-consumatore nel Capitolo 5; il contenuto normale nel Capitolo 6. Il problema del «livello di descrizione» dei sistemi è discusso nella sezione 6,3.

Millikan, Ruth G. (1993). *White Queen Psychology and Other Essays for Alice*. Cambridge, MA: MIT Press. — Raccolta di saggi che sviluppano e difendono la teleosemantica contro le obiezioni principali. Il saggio «Biosemantics» (cap. 6) risponde al problema della mismisrappresentazione sistematica; «Compare and Contrast Dretske, Fodor, and Millikan on Teleosemantics» (cap. 7) è una rassegna critica delle teorie naturalistiche.

V. Semantica cognitiva e linguistica cognitiva

Rosch, Eleanor (1973). «Natural Categories». *Cognitive Psychology*, 4(3), 328–350. — L'articolo fondativo della teoria dei prototipi. La struttura dei giudizi di tipicità per categorie naturali è presentata negli Esperimenti 1–4; la nozione di «somiglianza familiare» wittgensteiniana è messa in relazione con i dati nella sezione 4.

Rosch, Eleanor, Mervis, Carolyn B. (1975). «Family Resemblances: Studies in the Internal Structure of Categories». *Cognitive Psychology*, 7(4), 573–605. — La formalizzazione della somiglianza familiare come struttura delle categorie naturali. Il «cue validity» come misura della diagnosticità delle caratteristiche prototipiche è definito nella sezione 2.

Rosch, Eleanor, Mervis, Carolyn B., Gray, Wayne, Johnson, David, Boyes-Braem, Penny (1976). «Basic Objects in Natural Categories». *Cognitive Psychology*, 8(3), 382–439. — L'identificazione empirica dei livelli di base come livelli di massima distintività tra categorie e massima coesione interna. I criteri convergenti (similarità percettiva, generalizzabilità motoria, lessicale) sono nel paragrafo 3. Rilevante per la sezione X.3,4 e XII.8,1.

Fillmore, Charles J. (1975). «An Alternative to Checklist Theories of Meaning». In: Cogen, C., et al. (eds.), *Proceedings of the First Annual Meeting of the Berkeley Linguistics Society*. Berkeley: BLS, 123–131. — L'introduzione della nozione di «scena» come precursore della semantica dei frame. La critica della semantica delle condizioni necessarie e sufficienti è nella sezione 2.

Fillmore, Charles J. (1982). «Frame Semantics». In: *Linguistics in the Morning Calm*. Seoul: Hanshin Publishing Co., 111–137. — La formulazione sistematica della semantica dei frame. La struttura del frame della vendita come esempio paradigmatico è nel paragrafo 3; la profilazione come selezione all'interno del frame è nel paragrafo 5. Rilevante per la sezione X.2.

Fillmore, Charles J., Atkins, Beryl T. S. (1992). «Toward a Frame-Based Lexicon: The Semantics of RISK and Its Neighbors». In: Lehrer, A., Kittay, E. F. (eds.), *Frames, Fields, and Contrasts*. Hillsdale: LEA, 75–102. — Applicazione della semantica dei frame al campo semantico del rischio, con analisi dettagliata delle relazioni tra profilazione e struttura argomentale.

Lakoff, George (1987). *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. Chicago: University of Chicago Press. Trad. it.: (1987). *Donne, fuoco e cose pericolose*. Milano: Melangolo. — La proposta della semantica cognitiva radicale anti-oggettivista. La struttura radiale delle categorie è nel Capitolo 6; gli schemi di immagine nel Capitolo 17;

la proiezione metaforica come morfismo tra domani nel Capitolo 11. Rilevante per la sezione X.3 e X.4.

Lakoff, George, Johnson, Mark (1980). *Metaphors We Live By*. Chicago: University of Chicago Press. Trad. it.: (1998). *Metafora e vita quotidiana*. Milano: Bompiani. — L'analisi sistematica delle metafore concettuali come strutture che organizzano il pensiero e il linguaggio. Lo schema CONTENITORE come metafora di base è nel Capitolo 3; la metafora strutturale e quella orientazionale nei Capitoli 4–5. Rilevante per la sezione X.4,2.

Barsalou, Lawrence W. (1987). «The Instability of Graded Structure: Implications for the Nature of Concepts». In: Neisser, U. (ed.), *Concepts and Conceptual Development*. Cambridge: Cambridge University Press, 101–140. — La prova empirica dell'instabilità del grado di tipicità tra contesti e tra occasioni, con la proposta dei «concetti situati» come alternativa ai prototipi stabili. Rilevante per le sezioni X.3,2 e XII.3,1.

Barsalou, Lawrence W. (1999). «Perceptual Symbol Systems». *Behavioral and Brain Sciences*, 22(4), 577–609. — La proposta teorica dei sistemi di simboli percettivi come base del contenuto concettuale. La nozione di simulatore come struttura che media l'attivazione contestuale dei concetti è nel paragrafo 3,2.

Tversky, Amos (1977). «Features of Similarity». *Psychological Review*, 84(4), 327–352. — La prova formale che la similarità non è una relazione simmetrica né transitiva in generale, con implicazioni per qualsiasi teoria della categorizzazione basata sulla distanza da un prototipo. Il Teorema 1 (asimmetria) e il Teorema 3 (non-transitività) sono i risultati rilevanti per la critica nella sezione XII.3.

VI. Modelli computazionali della categorizzazione

Nosofsky, Robert M. (1986). «Attention, Similarity, and the Identification-Categorization Relationship». *Journal of Experimental Psychology: General*, 115(1), 39–57. — L'articolo fondativo del Generalized Context Model (GCM). La formula della probabilità di classificazione (equazione 5) e la funzione di similarità esponenziale sono introdotte nella sezione 2. La relazione con il prototipo come caso limite è discussa nella sezione 4.

Ashby, F. Gregory, Maddox, W. Todd (1993). «Relations Between Prototype, Exemplar, and Decision Bound Models of Categorization». *Journal of Mathematical Psychology*, 37(3), 372–400. — Confronto sistematico tra modelli basati sul prototipo, sugli esemplari e sulle

decision boundaries. La prova che il modello GRT (General Recognition Theory) generalizza il GCM con distribuzioni gaussiane è nel Teorema 2. Rilevante per la sezione XII.4.2.

Kruschke, John K. (1992). «ALCOVE: An Exemplar-Based Connectionist Model of Category Learning». *Psychological Review*, 99(1), 22–44. — Il modello ALCOVE come integrazione del GCM con una rete neurale attenzionale. La derivazione della regola di apprendimento degli attenzioni è nel Appendice A. Rilevante come precedente computazionale per il GCM-firma della sezione XII.5.

Nosofsky, Robert M. (1992). «Similarity Scaling and Cognitive Process Models». *Annual Review of Psychology*, 43, 25–53. — Rassegna critica dei modelli computazionali della categorizzazione. La discussione della scelta dello spazio delle caratteristiche come problema fondamentale per qualsiasi modello basato sulla similarità è nella sezione 3.

VII. Neuroscienze della cognizione e dei sistemi di rappresentazione

Kanwisher, Nancy, McDermott, Josh, Chun, Marvin M. (1997). «The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception». *Journal of Neuroscience*, 17(11), 4302–4311. — L'evidenza fMRI originale per la FFA come area specializzata nel riconoscimento dei volti. Rilevante per la sezione III.4.2 in quanto evidenza contro l'incapsulamento rigido delle aree visive.

Haxby, James V., Gobbini, M. Ida, Furey, Maura L., Ishai, Alomit, Schouten, Jennifer L., Pietrini, Pietro (2001). «Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex». *Science*, 293(5539), 2425–2430. — La prova che il riconoscimento dei volti non è localizzato esclusivamente nella FFA ma distribuito in pattern di attivazione che coinvolgono aree multiple. Evidenza empirica a supporto della tesi della ridondanza costitutiva.

Quiroga, Rodrigo Quian, Reddy, Leila, Kreiman, Gabriel, Koch, Christof, Fried, Itzhak (2005). «Invariant Visual Representation by Single Neurons in the Human Brain». *Nature*, 435(7045), 1102–1107. — La scoperta dei «neuroni di Jennifer Aniston» nell'ippocampo umano come esempio di rappresentazione sparse ad alta invarianza. Rilevante per la discussione sulla tensione tra rappresentazioni sparse e ridondanza distribuita nei sistemi reali.

Friston, Karl J. (2010). «The Free-Energy Principle: A Unified Brain Theory?». *Nature Reviews Neuroscience*, 11(2), 127–138. — La formulazione sistematica del principio dell'energia libera come teoria unificata della

percezione, dell'azione e dell'apprendimento. La connessione con la minimizzazione della sorpresa bayesiana è nel paragrafo 2; la struttura gerarchica del modello predittivo nel paragrafo 3. Rilevante per la sezione III.5 e IV.6.

Friston, Karl J., et al. (2017). «Active Inference and Epistemic Value». *Cognitive Neuroscience*, 8(4), 187–225. — L'estensione del principio dell'energia libera all'inferenza attiva e alla spiegazione dell'esplorazione. Rilevante per la sezione sulle rappresentazioni procedurali nella firma informazionale.

Singer, Wolf (1999). «Neuronal Synchrony: A Versatile Code for the Definition of Relations?». *Neuron*, 24(1), 49–65. — L'ipotesi della sincronizzazione oscillatoria gamma come meccanismo di binding percettivo. La difficoltà che questo pone per le teorie del portatore localizzato è il punto di partenza della sezione III.2.2.3.

Dehaene, Stanislas, Changeux, Jean-Pierre (2011). «Experimental and Theoretical Approaches to Conscious Access». *Neuron*, 70(2), 200–227. — La teoria dello spazio di lavoro globale come framework per la coscienza d'accesso. La distinzione tra processamento locale e processamento globale nella diffusione dell'informazione rilevante per la sezione IV sull'architettura della ridondanza gerarchica.

Mediano, Pedro A. M., Rosas, Fernando E., Farah, Jonathan C., Shanahan, Murray, Bor, Daniel, Seth, Anil K., Barrett, Adam B. (2022). «Greater than the Parts: A Review of the Information Decomposition Approach to Consciousness». *Philosophical Transactions of the Royal Society B*, 377, 20210059. — Rassegna recente dell'applicazione della PID e dell'O-information alla neuroscienza della coscienza. Discute sia le promesse sia i limiti computazionali degli approcci informativi. Rilevante per la sezione VIII del testo.

VIII. Psicopatologia e neuropsicologia della cognizione

Sacks, Oliver (1985). *The Man Who Mistook His Wife for a Hat*. New York: Summit Books. Trad. it.: (1986). *L'uomo che scambiò sua moglie per un cappello*. Milano: Adelphi. — Casi clinici che illustrano le dissociazioni tra diversi sistemi di rappresentazione, in particolare la prosopagnosia associativa (Capitolo 1) e le sindromi di asimbolie (Capitolo 3). Usato come fonte di casi fenomenologici nella sezione II.5.

Young, Andrew W., et al. (1993). «Configurational Information in Face Perception». *Perception*, 22(2), 195–215. — La dimostrazione sperimentale che il riconoscimento dei volti dipende dalla configurazione globale

(relazioni tra le parti) più che dalle caratteristiche locali, con evidenza da pazienti con prosopagnosia. Rilevante per l'analisi della firma del riconoscimento del volto.

Ellis, Hadyn D., Young, Andrew W. (1990). «Accounting for Delusional Misidentifications». *British Journal of Psychiatry*, 157, 239–248. — La teoria del «doppio deficit» per la sindrome di Capgras (lesione al circuito di riconoscimento familiare più pensiero delirante). Rilevante per la sezione II sulla classificazione delle varietà di ridondanza patologica.

Ramachandran, Vilayanur S., Blakeslee, Sandra (1998). *Phantoms in the Brain*. New York: William Morrow. Trad. it.: (1999). *Fantasma nel cervello*. Milano: Mondadori. — L'analisi degli arti fantasma e di altri fenomeni neuropsicologici come evidence per la struttura distribuita dello schema corporeo. Rilevante per la convergenza con la fenomenologia merleau-pontiana nella sezione IX.4.

IX. Intelligenza artificiale, apprendimento automatico e semantica vettoriale

Mikolov, Tomáš, Chen, Kai, Corrado, Greg, Dean, Jeffrey (2013). «Efficient Estimation of Word Representations in Vector Space». *arXiv:1301.3781*. — L'introduzione di word2vec e delle rappresentazioni vettoriali dense per le parole. La struttura delle dipendenze contestuali nei vettori è il caso di studio principale della sezione VIII sui sistemi artificiali.

Vaswani, Ashish, et al. (2017). «Attention Is All You Need». In: *Advances in Neural Information Processing Systems 30*. — L'architettura Transformer e il meccanismo di self-attention come implementazione di una forma di ridondanza gerarchica. La connessione tra l'attention multi-head e la struttura di ridondanza della firma è discussa nella sezione VIII.3.

Belghazi, Mohamed Ishmael, et al. (2018). «MINE: Mutual Information Neural Estimation». In: *Proceedings of ICML 2018*, 531–540. — Il metodo di stima variazionale dell'informazione mutua attraverso reti neurali. La garanzia di convergenza (Teorema 1) e il bias dello stimatore per campioni finiti sono rilevanti per la sezione IV.5bis.3 sulle strategie di approssimazione della PID.

Marcus, Gary, Davis, Ernest (2019). *Rebooting AI: Building Artificial Intelligence We Can Trust*. New York: Pantheon. Trad. it.: (2020). *Reboot AI*. Torino: Einaudi. — Critica ampia delle limitazioni dei sistemi di apprendimento automatico basati su corpus, con enfasi sul problema del grounding e della comprensione. Il capitolo sulla «debolezza della

comprensione» è il contesto del confronto tra firma genuina e firma simulata nella sezione VIII.

X. Filosofia della mente: questioni fondazionali

Dennett, Daniel C. (1987). *The Intentional Stance*. Cambridge, MA: MIT Press. Trad. it.: (1993). *L'atteggiamento intenzionale*. Bologna: Il Mulino. — La proposta dello «stance intenzionale» come strategia predittiva del comportamento. Il confronto con la tesi del portatore fisico del contenuto è in «True Believers» (cap. 2).

Putnam, Hilary (1975). «The Meaning of "Meaning"». In: Gunderson, K. (ed.), *Language, Mind, and Knowledge*. Minneapolis: University of Minnesota Press, 131–193. Ristampato in: Putnam, H. (1975). *Mind, Language and Reality*. Cambridge: Cambridge University Press. — L'argomento della Terra Gemella per l'esternalismo semantico. Il caso dell'acqua come H₂O versus XYZ è rilevante per la discussione del grado di oggettività del contenuto nella sezione V.5.

Burge, Tyler (1979). «Individualism and the Mental». *Midwest Studies in Philosophy*, 4(1), 73–121. — L'argomento dell'artrite (pensieri da individuare in parte dal contesto sociale) come secondo pilastro dell'esternalismo. La questione di come la firma informazionale catturi l'esternalismo senza presupporre un portatore localizzato è affrontata nelle sezioni I e IX.

Chalmers, David J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press. Trad. it.: (1997). *La mente cosciente*. Milano: McGraw-Hill. — La distinzione tra problemi «facili» e problema «difficile» della coscienza. Il problema difficile è lasciato fuori dallo scopo della presente monografia (nota I.1), ma la discussione di Chalmers sulle rappresentazioni funzionali è rilevante per la sezione sulla normatività implicita.

Merleau-Ponty, Maurice (1945). *Phénoménologie de la perception*. Paris: Gallimard. Trad. it.: (2003). *Fenomenologia della percezione*. Milano: Bompiani. — La fenomenologia dello schema corporeo come struttura che integra componenti motorie, propriocettive e visive. Il Capitolo III («Il corpo proprio») sviluppa la nozione di intenzionalità motoria. Rilevante per la sezione IX.4 e per la connessione tra firma procedurale e schema corporeo.

Dewey, John (1929). *Experience and Nature*. Second edition. La Salle: Open Court. Trad. it. parziale: (1990). *Esperienza e natura*. Milano: Mursia. — La critica pragmatista alla dicotomia interno/esterno nel pensiero. La

nozione di «transazione» tra organismo e ambiente è il precursore filosofico della nozione di firma come struttura relazionale tra sistema e mondo.

XI. Studi di revisione e rassegne critiche

Piantadosi, Steven T. (2021). «Meaning Without Reference in Large Language Models». *Notes*, 55(4), 831–856. — Argomentazione filosofica che i sistemi linguistici su larga scala possono avere «significato funzionale» senza riferimento causale al mondo, con implicazioni per la distinzione firma genuina/simulata.

Tishby, Naftali, Schwartz, Elie, Bialek, William (1999). «The Information Bottleneck Method». In: *Proceedings of the 37th Allerton Conference on Communication, Control, and Computing*, 368–377. — Il metodo dell'information bottleneck come principio di compressione ottimale. La relazione con la PID — entrambi gli approcci cercano di isolare la struttura informazionale «rilevante» rispetto a un obiettivo — è discussa nella sezione IV.1.

Schneidman, Elad, Berry, Michael J., Segev, Ronen, Bialek, William (2006). «Weak Pairwise Correlations Imply Strongly Correlated Network States in a Neural Population». *Nature*, 440(7087), 1007–1012. — La dimostrazione che la struttura di dipendenza statistica nelle reti neurali non è catturata dalle correlazioni a coppie ma richiede misure di ordine superiore. Evidenza empirica per la sinergia di ordine superiore nei sistemi neurali.

Tononi, Giulio (2004). «An Information Integration Theory of Consciousness». *BMC Neuroscience*, 5, 42. — La teoria Φ dell'integrazione dell'informazione. La relazione tra Φ e la firma informazionale è quella tra una misura scalare globale (Φ) e una struttura reticolare dettagliata (la firma): Φ cattura un aspetto importante ma non l'intera struttura.

XII. Opere di consultazione matematica

Gromov, Mikhail (1981). «Structures métriques pour les variétés Riemanniennes». In: Lafontaine, J., Pansu, P. (eds.), *Structures métriques pour les variétés riemanniennes*. Paris: CEDIC/Fernand Nathan. — La fonte originale della distanza di Gromov-Hausdorff per varietà riemanniane, poi generalizzata agli spazi metrici astratti da Gromov stesso e da Burago-Burago-Ivanov. La definizione come infimo su immersioni isometriche è nella sezione 3,2.

Burago, Dmitri, Burago, Yuri, Ivanov, Sergei (2001). *A Course in Metric Geometry*. Providence: AMS. — Il trattamento sistematico della distanza di Gromov-Hausdorff (Capitolo 7) con proprietà metriche, spazi di Gromov precompatti, e applicazioni alla geometria riemanniana. La prova che d_{GH} è una distanza (zero sse isometria, simmetria, disuguaglianza triangolare) è nel Teorema 7.3.30.

Sturm, Karl-Theodor (2003). «Probability Measures on Metric Spaces of Nonpositive Curvature». In: Auscher, P., et al. (eds.), *Heat Kernels and Analysis on Manifolds, Graphs, and Metric Spaces*, 357–390. — L'esistenza e l'unicità della media di Fréchet in spazi $CAT(0)$ (Teorema 4,3). Rilevante per la sezione A.17 sull'esistenza del punto fisso della mappa F .

Kraskov, Alexander, Stögbauer, Harald, Grassberger, Peter (2004). «Estimating Mutual Information». *Physical Review E*, 69(6), 066138. — La derivazione dello stimatore KSG dell'informazione mutua con la prova di consistenza e la stima del bias (equazione 8 e seguenti). Rilevante per il protocollo neuroimaging della sezione A.21.

Vigna, Sebastiano (2008). «A Broader Picture of the Dedekind Numbers». *Journal of Integer Sequences*, 11, 08.1.2. — Rassegna delle proprietà dei numeri di Dedekind $D(n)$ con la formula asintotica e la relazione con i coefficienti binomiali centrali. Rilevante per la sezione A.20.